EDV in Medizin und Biologie **10** (3), S. 65–69, ISSN 0300-8282

Der klinische Befund in der Sicht formaler Sprachen

von Hans-Jürgen Seelos*

Zusammenfassung

In der vorliegenden Arbeit wird versucht, die bei der Behandlung formaler Sprachen gewonnenen Erkenntnisse in Analogie zu den bekannten Programmiersprachen für die Generierung einer Befundsprache zu verwenden. Dabei wird insbesondere auf deren Bedeutung für die Synthese und Analyse von Befunden hingewiesen.

Summary

In the present paper the author tries to apply the knowledge obtained with the formal languages in order to create a clinical-finding language in analogy to the usual programming languages. Special consideration is given to its importance for the synthesis and analysis of clinical-findings.

1 Einleitung

»Das Problem der on-line Erfassung ›harter‹ Daten, etwa im Bereich der Labor- oder Biosignalverarbeitung, kann heute trotz mancher Schwierigkeiten als gelöst betrachtet werden. Im Gegensatz dazu ist die Erfassung jener Informationen, die das Ergebnis direkter ärztlicher Tätigkeit sind, noch weit von einer befriedigenden Lösung entfernt« [9].

Als Konsequenz daraus ergibt sich die Aufgabe, neue Methoden zu entwickeln und diese anzuwenden. So ist es nicht verwunderlich, daß die Verfügbarkeit der Methoden der Informatik in der Medizin [7, 12] auch dazu geführt hat, die Bemühungen um eine syntaktische und semantische Standardisierung der medizinischen Terminologie zu intensivieren [13, 14, 15, 17, 18].

Im folgenden soll, als weiterer Beitrag zu diesem Problemkreis, die Fragestellung untersucht werden, ob die medizinische Fachsprache im Sinne einer formalen Sprache definiert werden kann und welche Konsequenzen sich gegebenenfalls daraus ableiten lassen.

2 Methode

Zur Beschreibung der medizinischen Fachsprache bietet es sich an, die Semiotik heranzuziehen. Die Semiotik, d. h. die Lehre von den Zeichen und ihrer Anordnung, unterscheidet im einzelnen drei Ebenen:

- Syntax
  Die Regeln einer Grammatik, welche die Aneinanderreihung von Wörtern definieren, bezeichnet man als Syntax. Z. B. <SATZ> = <SUBJEKT> <PRAEDIKAT> <OBJEKT>. Syntaktisch falsch wäre daher der Satz »Die Nieren das Blut reinigen«.
- Semantik
  Als Semantik bezeichnet man den Teil einer Grammatik, der sich mit der Bedeutung der Wörter und Sätze beschäftigt.
  So ist etwa der Satz »Das Blut ist schwerhörig« syntaktisch korrekt, während ihm aber keine semantisch brauchbare Interpretation zukommt.
- Pragmatik
  Die von Umständen, wie Zeit, Gefühlen und subjektiven Einstellungen, abhängigen Einflüsse bei der Verwendung von Zeichen bezeichnet man als Pragmatik.
  Beispielsweise wird die kommunikative Beziehung beim ärztlichen Gespräch weitgehend vom ersten Eindruck des Arztes vom Patienten (Geschlecht, Lebensalter, Haarfarbe, Kleidung, allgemeine Erscheinung) bestimmt, der sich mit Voreinstellungen sowohl auf seiten des Arztes als auch auf seiten des Patienten überlagern kann.

Versteht man die medizinische Fachsprache als Untermenge einer natürlichen Sprache, wie z. B. Deutsch, so können ihre Eigenschaften folgendermaßen charakterisiert werden. Gegeben ist eine Menge von Vokabeln, die zur Spezifikation eines Befundes erforderlich sind (Befundbeschreibung). Mit diesem Vokabular kann man Sätze bilden (Befund). Dazu müssen die Vokabeln entsprechend einem definierten Regelsystem syntaktisch richtig angeordnet sein. Für eine medizinische Aussage (z. B. Befund) ist aber die syntaktische Korrektheit nur eine notwendige, aber keine hinreichende Bedingung. Zusätzlich gilt es auch die semantische Korrektheit zu gewährleisten. Die nachstehenden Ausführungen sollen sich daher nur auf die syntaktische und semantische Korrektheit einer medizinischen Aussage beschränken, zumal diese für die Handhabung der Pragmatik Voraussetzung sind.

Läßt sich die Syntax relativ leicht einer formalen Beschreibung zugänglich machen, so bereitet dies bei der Semantik in der Regel Schwierigkeiten. Wie dies aber für ein eng begrenztes Vokabular trotzdem möglich ist, soll in den nächsten Abschnitten am Beispiel einer Chomsky-Sprache zur Leber-Befundung und zur Befundung einer Röntgenbildaufnahme des Thorax aufgezeigt werden.

Eine Sprache, deren Syntax gewissen formalen Regeln genügt und deren Grammatik formalisierbar oder auch mathematisierbar ist, bezeichnet man als formale Sprache.

Ausgangspunkt für die Generierung einer Befundsprache ist der von Chomsky im Jahre 1959 eingeführte Begriff der Satzgliederungssprache [3, 4]. Es sei zunächst der Begriff »Satzgliederungsgrammatik« eingeführt.

Eine Satzgliederungsgrammatik G besteht aus einem Quadrupel $G = (V_T, V_N, P, S)$

*) Arbeitsgemeinschaft für Gemeinschaftsaufgaben der Krankenversicherung, Rellinghauser Straße 93–95, 4300 Essen 1

mit V_T = Menge (Alphabet) der Terminals
Terminals sind Zeichen, aus denen die Objekte einer Sprache bestehen.

V_N = Menge (Alphabet) der Nonterminals
Nonterminals sind Begriffe, mit deren Hilfe die Syntax beschrieben wird. Nonterminals werden ausschließlich mit Großbuchstaben bezeichnet z. B. »ORGAN«.

$S \varepsilon V_N$ ist ein ausgezeichnetes Nonterminal mit der syntaktischen Bedeutung »Satz«.

P: ist ein Regelsystem mit endlich vielen Regeln der Form $\varphi_i \rightarrow \omega_i$ ($i = 1, 2, \ldots, n$) wobei φ_i und ω_i Strings über V mit $V = V_N \vee V_T$ sind.

Für eine zur Satzgliederungsgrammatik G gehörige Satzgliederungssprache L(G) gilt somit: Ist $W(V_T)$ die Menge aller Sätze über V_T, dann besteht L(G) aus allen denjenigen Sätzen $q \varepsilon W(V_T)$, die entsprechend dem Regelsystem P aus dem ausgezeichneten Nonterminal S ableitbar sind:

$L(G) = \{q | q \varepsilon W(V_T) \wedge S \Rightarrow P\}$.

In Abhängigkeit der Einschränkungen des Regelsystems P unterscheidet man verschiedene Typen von Satzgliederungssprachen:

- C_3 (reguläre Sprachen),
- C_2 (kontextfreie Sprachen),
- C_1 (kontextsensitive Sprachen) und
- C_0 (allgemeine Sprachen).

Ist C_i ($i = 0, 1, 2, 3$) die Menge der Satzgliederungssprachen vom Typ i, so gilt: $C_3 \subset C_2 \subset C_1 \subset C_0$.

Für den Anwendungsbereich Befundsprache sollen im weiteren die Typen C_3 und C_2 diskutiert werden, bieten diese trotz ihres relativ einfachen Regelsystems die Möglichkeit zur formalen Beschreibung medizinischer Befunde, wie die nachfolgenden Beispiele demonstrieren. Hierbei umfaßt das Alphabet der Terminals die Vokabeln der Befundsprache. Dem ausgezeichneten Nonterminal S, kommt die syntaktische Bedeutung »Befundaussage« zu.

Reguläre Sprachen (Typ C_3)

Sind die Regeln des Regelsystems P einer Grammatik alle von der Form $A \rightarrow a|aB$ bzw. $A \rightarrow a|Ba$, so heißt die Satzgliederungssprache vom Typ C_3 (dabei gilt $a \varepsilon V_T$, $A,B \varepsilon V_N$)

Beispiel 1:

Gesucht sei eine Satzgliederungsprache L(G) vom Typ C_3 zur Beschreibung eines Leberbefundes mit

$L(G) = \{q \mid q$ beschreibt einen Leberbefund mit den Parametern Druckschmerz, Konsistenz, Oberfläche, Rand, Größe$\}$.

Ein Befund, d. h. die verbale Darstellung eines medizinischen Tatbestandes, hat zu nachstehenden Fragen Stellung zu nehmen

- was (Organ),
- wo (Topographie, Lokalisation),
- wie (Morphologie)

verändert resp. unverändert (Nullbefund [16]) ist. Ein Leberbefund ist demnach durch die Angabe des Organs (Leber) und seiner Morphologie ausreichend charakterisiert. Letztgenannte möge durch die Parameter Druckschmerz, Konsistenz, Oberfläche, Rand und Größe beschrieben werden. Die entsprechenden Ausprägungen lauten somit für

- Druckschmerz: vorhanden, oB
- Konsistenz: weich, derb, hart
- Oberfläche: fein, grob, knotig
- Rand: stumpf, scharf, gekerbt
- Größe: vergrößert.

Bildet man alle möglichen semantisch korrekten Kombinationen zwischen den aufgeführten Parametern und ihren entsprechenden Ausprägungen in ein Regelsystem ab, so kann die zu L(G) gehörige Satzgliederungsgrammatik $G = (V_T, V_N, P, S)$ angegeben werden:

V_T = {Leber, Druckschmerz, Konsistenz, Oberfläche, Rand, vergrößert, vorhanden, oB, weich, derb, hart, fein, grob, knotig, stumpf, scharf, gekerbt}

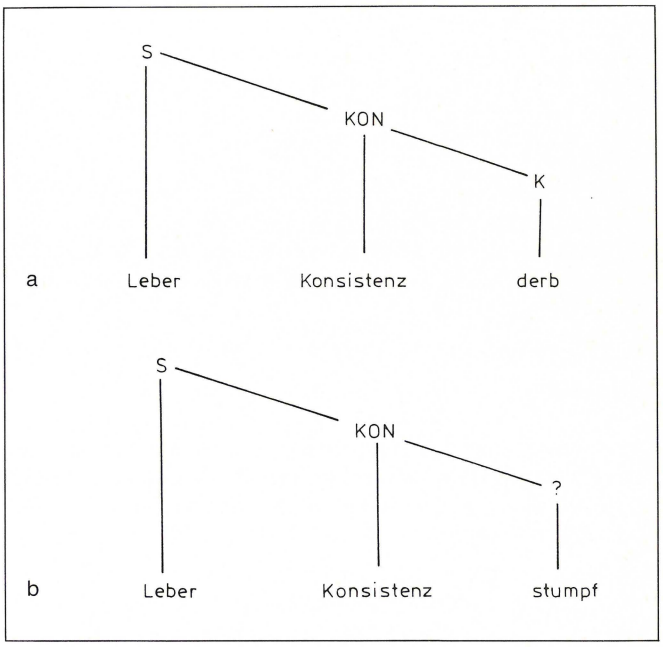
V_N = {DRU, KON, OBER, RAN, GRO, X, K, O, R, S}

$S \varepsilon V_N$

P:
S → Leber DRU | Leber KON | Leber OBER | Leber RAN | Leber GRO
DRU → Druckschmerz X
X → vorhanden | oB
KON → Konsistenz K
K → weich | derb | hart
OBER → Oberfläche 0
O → fein | grob | knotig
RAN → Rand R
R → stumpf | scharf | gekerbt
GRO → vergrößert

Da ausschließlich Regeln der Form $A \rightarrow a\,B$ resp. $A \rightarrow a$ ($A, B \varepsilon V_N$, $a \varepsilon V_T$) verwendet wurden, bezeichnet man G und die zugehörige Sprache L(G) als rechtslinear. Dieses Regelsystem genügt nicht nur der eingangs aufgezeigten Syntax eines klinischen Befundes, sondern berücksichtigt auch den semantischen Aspekt, wie die Syntaxanalyse zeigt.

Abb. 1. Syntaxbaum zu der Grammatik der C_3-Leber-Sprache (nach dem Top-down-Verfahren).
a korrekter Eingabestring
b unkorrekter Eingabestring



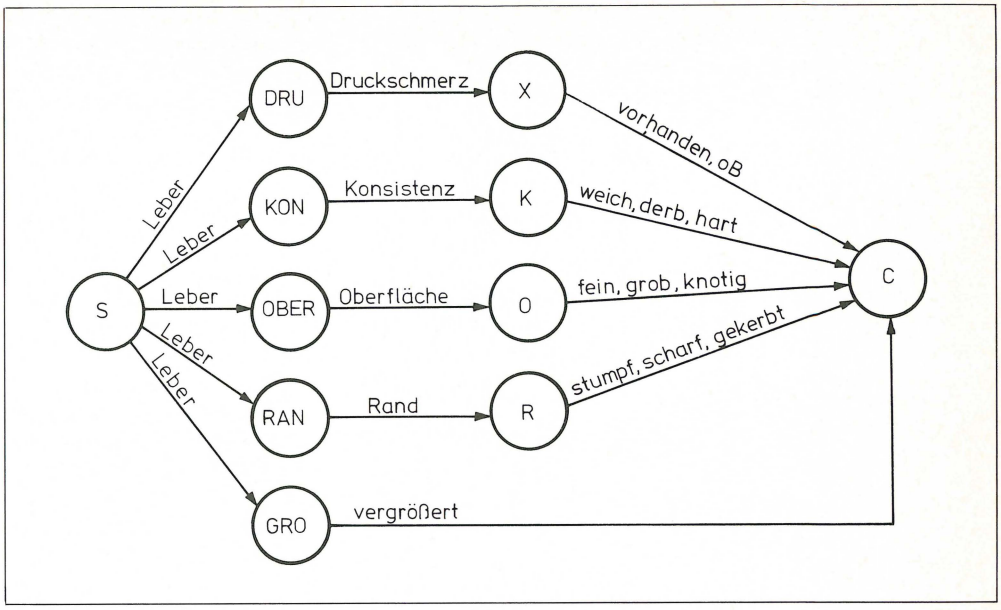


Abb. 2. Übergangsgraph des nichtdeterminierten endlichen Automaten zu der Grammatik der C_3-Leber-Sprache mit Startzustand S und Endzustand C. Jede Befundaussage der Satzgliederungssprache L(G) = {q | q beschreibt einen Leberbefund mit den Parametern Druckschmerz, Konsistenz, Oberfläche, Rand, Größe} entspricht einem Weg durch den Übergangsgraphen des Automaten π.

Die Syntaxanalyse hat zum Ziel, möglichst effektiv festzustellen, ob ein gegebener Terminalstring x zur Sprache gehört oder nicht. Die Aufgabe besteht in der Konstruktion eines Syntaxbaumes. Dabei kann einmal durch Rückverfolgung der Produktionsregeln ausgehend von den Terminals des Strings x das ausgezeichnete Nonterminal S erreicht werden (Bottom-up-Verfahren), andererseits kann man aber auch den Syntaxbaum von S ausgehend durch Anwendung der Produktionsregeln konstruieren und versuchen, die Terminals von x zu erreichen (Top-down-Verfahren) [6].

Durch die Formalisierung der Befundsprache schlägt sich ein Teil der semantischen Unrichtigkeiten in der Syntax nieder. Sie lassen sich dann durch eine Syntaxprüfung ermitteln. So würde beispielsweise der semantisch unkorrekte String »Leber Konsistenz stumpf« bei einer Syntaxanalyse im Gegensatz zum korrekten String »Leber Konsistenz derb« (Abb. 1 a) eine Fehlermeldung provozieren (Abb. 1 b). Algorithmen für entsprechende Computer-Programme zur Syntaxprüfung, Parser genannt, finden sich etwa bei FOSTER [6] und HERSCHEL [8], so daß an dieser Stelle auf deren Darstellung nicht näher eingegangen werden soll. Jedoch möchte der Verfasser darauf hinweisen, daß sich mit solchen Verfahren zur Syntaxanalyse eine elegante Möglichkeit zur Prüfung von Befunden im Klartext andeutet. Wird der Befunder beispielsweise durch vorgegebene Strukturen auf einem Erfassungsformular dazu angehalten, Freitextangaben im Sinne einer Befundsprache zu spezifizieren, kann vor der rechnergestützten Erstellung eines Befundberichtes [9, 14, 15] der Freitext auch auf semantische Korrektheit überprüft werden.

Äquivalent zu der Grammatik der C_3-Leber-Sprache kann bei regulären Sprachen ein endlicher Automat $\pi = (A, B, Z, \gamma)$

wobei A = Eingabealphabet, B = Ausgabealphabet,
Z = Zustandsmenge, γ = Abbildungsfunktion

definiert werden.

Der Zusammenhang zwischen rechtslinearen Grammatiken und endlichen Automaten ergibt sich aus nachstehendem Verfahren [8]. Hierbei gilt U, V, C ε V_N und n, b ε V_T.

(i) Jedem Nonterminal entspricht ein Knoten (Zustand), jedem Zustand ein Nonterminal. Der Startzustand entspricht dem ausgezeichneten Nonterminal S.

(ii) Jeder Regel der Form U → nV entspricht eine Kante vom Knoten U zum Knoten V mit der Bewertung n.

(iii) Jeder Regel U → b entspricht eine Kante vom Knoten U zum Knoten C, wobei für den Endzustand das neue Nonterminal C eingeführt wird.

Ein Terminalstring x ist genau dann richtig, wenn der Automat π bei der Abarbeitung des Strings x vom Start-(S) in den Endzustand (C) übergeht.

Wie aus der Abb. 2 ersichtlich ist, akzeptiert der Automat π ausschließlich semantisch und syntaktisch korrekte Terminalstrings als »Eingabeworte«. Dies wurde dadurch erreicht, daß alle semantisch korrekten Terminalkombinationen mit dem Regelsystem P abgebildet werden konnten. Als alternative Darstellungsform des Übergangsgraphen (Abb. 2) kann die Übergangsmatrix (Tab. 1) und die Abbildungsfunktion angegeben werden [8].

Im Gegensatz zum oben angeführten Beispiel 1 einer C_3-Leber-Sprache muß die Syntax der meisten Programmier- und Befundsprachen durch eine C_2-Sprache beschrieben werden, weil dieser Sprachtyp eine Erweiterung des Regelsystems P zuläßt. Eine Satzgliederungssprache L(G) heißt vom Typ 2 oder kontextfrei, wenn alle Regeln von der Form A → ω sind (ω ε V mit V = V_N ∨ V_T, A ε V_N). Die Bezeichnung kontextfrei bezieht sich auf den Umstand, daß eine Regel A → ω unabhängig vom Kontext angewendet werden kann.

Beispiel 2:

Die C_3-Leber-Sprache wird jetzt als C_2-Sprache definiert.

L(G) = {q | q beschreibt einen Leberbefund}
G = (V_T, V_N, P, S)
mit V_T = {Leber, Druckschmerz, Konsistenz, Oberfläche, Rand, vergrößert, vorhanden, oB, weich, derb, hart, fein, grob, knotig, stumpf, scharf, gekerbt}
V_N = {S, ORG, MOD, X, Y, Z, U, GRO}
S ε V_N
P: S → ORG MOD

MOD → Druckschmerz X | Konsistenz Y | Oberfläche Z | Rand U | GRO

Tab. 1. Übergangsmatrix für die Grammatik der C_3-Leber-Sprache

Zustände / Eingabealphabet	BEF	DR-U	KO-N	OB-ER	RA-N	GR-O	X	K	O	R	C
Leber	*	?	?	?	?	?	?	?	?	?	?
Druckschmerz	?	X	?	?	?	?	?	?	?	?	?
Konsistenz	?	?	K	?	?	?	?	?	?	?	?
Oberfläche	?	?	?	0	?	?	?	?	?	?	?
Rand	?	?	?	?	R	?	?	?	?	?	?
vergrößert	?	?	?	?	?	C	?	?	?	?	?
vorhanden	?	?	?	?	?	?	C	?	?	?	?
oB	?	?	?	?	?	?	C	?	?	?	?
weich	?	?	?	?	?	?	?	C	?	?	?
derb	?	?	?	?	?	?	?	C	?	?	?
hart	?	?	?	?	?	?	?	C	?	?	?
fein	?	?	?	?	?	?	?	?	C	?	?
grob	?	?	?	?	?	?	?	?	C	?	?
knotig	?	?	?	?	?	?	?	?	C	?	?
stumpf	?	?	?	?	?	?	?	?	?	C	?
scharf	?	?	?	?	?	?	?	?	?	C	?
gekerbt	?	?	?	?	?	?	?	?	?	C	?

* = DRU, KON, OBER, RAN, GRO
? = Fehlerausgang

Die Matrizenelemente beschreiben die Zustände, in die der Automat übergeht, wenn ein Eingabezeichen (Zeile der Matrix) im als Spaltenelement angegebenen Zustand eingegeben wird.

ORG →Leber
X →vorhanden | oB
Y →weich | derb | hart
Z →fein | grob | knotig
U →stumpf | scharf | gekerbt
GRO →vergrößert

Im Vergleich zur Grammatik des Beispiels 1 können die Regeln S → Leber DRU | Leber KON | Leber OBER| Leber RAN | Leber GRO
jetzt ersetzt werden durch

S →ORG MOD
ORG →Leber
MOD →Druckschmerz X | Konsistenz Y | Oberfläche Z | Rand U | GRO.

Das Regelsystem wurde also von ursprünglich 10 Regeln der Grammatik der C_3-Sprache auf nun 8 Regeln der Grammatik der C_2-Sprache reduziert. Durch die Zuordnung des Terminals »Leber« zum Nonterminal »ORG« ist jetzt bei komplexeren Anwendungsfällen die Möglichkeit gegeben, den Parameter Druckschmerz auch auf andere Abdominalorgane, wie etwa Milz und Nieren, zu beziehen.

Das Regelsystem P müßte dann lauten:

P: S →LI MOD | ORG MOD | ORG VER
LI →Niere | Milz
ORG →Leber
MOD →Druckschmerz X
X →vorhanden | oB
VER →Konsistenz Y | Oberfläche Z | Rand U | GRO
Y →weich | derb | hart
Z →fein | grob | knotig
U →stumpf | scharf | gekerbt
GRO →vergrößert

mit

V_T = {Niere, Leber, Milz, Druckschmerz, Konsistenz, Oberfläche, Rand, vergrößert, vorhanden, oB, weich, derb, hart, fein, grob, knotig, stumpf, scharf, gekerbt}
V_N = {S, MOD, ORG, VER, X, Y, Z, U, GRO, LI}

Das nachfolgende Beispiel 3 einer C_2-Sprache zur Röntgenbildbefundung einer Lungenaufnahme möge die bisherigen Ausführungen zusammenfassend veranschaulichen.

Beispiel 3:

L(G) = { q | q entspricht einem Röntgenbefund der Lunge mit den Parametern Verschattung, Kalkflecken, verminderte Gefäßzeichnung}
G = (V_N, V_T, P, S)
mit V_T = {Lunge, Verschattung, verminderte Gefäßzeichnung, Kalkflecken, mit Aufhellung, mit Ringschatten, Fremdkörper, kalkhart, dicht, wenig dicht, rundlich, eckig, generalisiert, streifig, flächig, rechts, links, beidseitig, oben, mitte, unten, unscharf, scharf}
V_N = {S, TOPO, MODI, Z, DICHT, FORM, R1, R2, BEG, ORG}
S ε V_N
P: S →ORG TOPO MODI
ORG →Lunge
TOPO →R1 | R1 R2
MODI →Verschattung Z | verminderte Gefäßzeichnung | Kalkflecken
Z →TOPO DICHT FORM BEG
DICHT →mit Aufhellung | mit Ringschatten Fremdkörper | kalkhart | dicht | wenig dicht
FORM →rundlich | eckig | generalisiert | streifig | flächig
R1 →rechts | links | beidseitig
R2 →oben | mitte | unten
BEG →unscharf | scharf

Wie unschwer einzusehen, handelt es sich um eine Sprache vom Typ C_2. Ein entsprechender Syntaxbaum ist in Abb. 3 dargestellt.

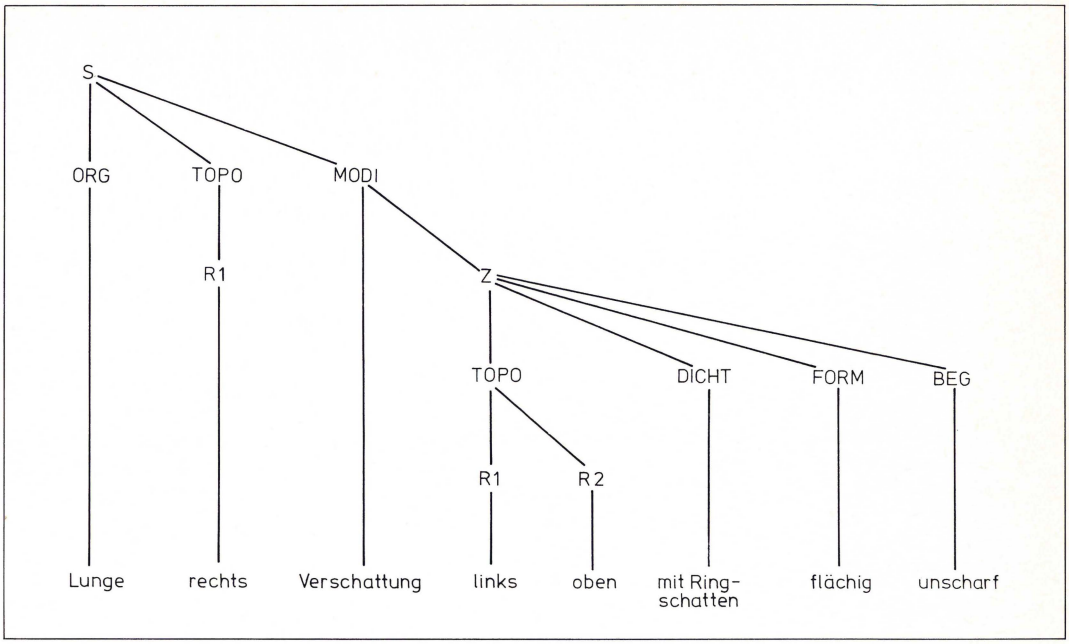


Abb. 3. Syntaxbaum für die Grammatik der C_2-Sprache des Beispiels 3 mit Eingabestring »Lunge rechts Verschattung links oben mit Ringschatten flächig unscharf«.

3 Ergebnis

An drei Beispielen wurde eine Methode aufgezeigt, mittels der Theorie formaler Sprachen eine »Befundsprache« zu entwickeln.

Die Grammatik solch einer Befundsprache kann von zwei unterschiedlichen Standpunkten aus betrachtet werden. Einmal kann sie als formale Methode zur Erzeugung aller möglichen richtigen Sätze einer Sprache aufgefaßt werden. Die Regeln sind dann zur Synthese aller zulässigen Befunde einer Sprache zu verwenden (Entwicklung von Befundbeschreibungen). In diesem Fall liegen die Schwierigkeiten darin, Grammatiken zur Erzeugung solcher Befundsprachen zu erstellen.

Andererseits kann ein vorliegender Befund syntaktisch und semantisch mittels Parsern (z. B. trial and error, bottom-up oder top-down Verfahren ohne Rücksetzen) geprüft werden. Dieser Aspekt ist insbesondere bei der Prüfung von klartextlich spezifizierten Befunden bedeutsam.

Darüber hinaus resultiert aus der theoretischen Durchleuchtung dieser Thematik ein Formalismus, der zwangsläufig eine höhere Präzision der dem Befunder angebotenen Befundbeschreibungen [5, 14] zur Folge haben wird.

Literaturverzeichnis

[1] BAR-HILLEL, Y.: Logical Syntax and Semantics. Language 30 (1954) 230–237
[2] BAR-HILLEL, Y., CARNAP, R.: Semantic Information. Brit. J. Phil. Sci. 4 (1953) 147–157
[3] CHOMSKY, N.: On Certain Formal Properties of Grammars. Inf. and Control 2 (1959) 137–16
[4] CHOMSKY, N., SCHÜTZENBERGER, M. P.: Algebraic Theory of Contextfree Languages. Proc. Blarium Symp. Programming Languages. Studies in Logic. North Holland 1962
[5] DETZEL, H., SEELOS, H.-J.: Befunderfassung bei der klinischen Untersuchung für die anschließende rechnergestützte Synthese von Befundberichten – Erste Erfahrungen. Der medizinische Sachverständige 3 (1979)
[6] FOSTER, J. M.: Automatische Syntax-Analyse (München: Hanser 1971)
[7] GALL, M. W.: Computer verändern die Medizin. Schriftenreihe der Bezirksärztekammer Nordwürttemberg. (Stuttgart: Genter 1969)
[8] HERSCHEL, R.: Einführung in die Theorie der Automaten, Sprachen und Algorithmen. (München-Wien: Oldenburg 1974)
[9] KOEPPE, P.: Zum Problem der EDV-gerechten Erfassung medizinischer Befunde. Method. Inform. Med. 1 (1971) 25–29
[10] MOYNE, J. A.: A Survey of Grammars and Recognizers for Formal and Natural Languages. Technical Report FSC 71-6006 IBM Fed. Sys. Div. Gaithersburg, Maryland (1971)
[11] PRATT, A. W.: Medicine, Computer and Linguistics. Adv. Biomed. Engineering 3 (1973) 97–240
[12] REICHERTZ, P. L.: Medizinische Informatik. IBM-Nachrichten 215 (1973)
[13] SCHNEIDER, W., SAGVALL HEIN, a.-l.: Computational linguistics in medicine. (Amsterdamm – New York – Oxford: North-Holland Publishing Company 1977)
[14] SEELOS, H.-J.: Inhaltliche Implementierung einer Befundbeschreibung für die allgemeinmedizinische außerklinische Untersuchung. Eine Vorstufe zur rechnergestützten Synthese von Befundberichten, dargestellt am Beispiel des Vertrauensärztlichen Dienstes. (Essen: Arbeitsgemeinschaft für Gemeinschaftsaufgaben der Krankenversicherung 1979)
[15] ECKER, F., KATZENBERGER, K., STIEBER, J., THURMAYR, R.: Habert-ein System zur computerunterstützten Röntgenbefundung. GSF-Bericht MD 114 (München: Gesellschaft für Strahlen- und Umweltforschung mbH 1975)
[16] WAGNER, G.: Bedeutung und Verläßlichkeit des Nullbefundes in der Medizin. Meth. Inform. Med. 5 (1966) 40–44
[17] WINGERT, F.: Morphosyntaktische Zerlegung von Komposita der medizinischen Sprache. Method. Inform. Med. 4 (1977) 248–255
[18] WINGERT, F.: Medical Language Data Processing. In: Reichertz, P. L., Goos, G. Informatics and Medicine (Berlin – Heidelberg – New York: Springer 1977)

Eingegangen: 11. 6. 1979
Anschrift des Verfassers: Dipl.-Inform. med. Hans-Jürgen Seelos, Arbeitsgemeinschaft für Gemeinschaftsaufgaben der Krankenversicherung, Rellinghauser Straße 93–95, 4300 Essen 1

EDV in Medizin und Biologie **10** (3), S. 70–79, ISSN 0300-8282

Automatic Generation of Design and Related Matrices by Formula Processing Computer Programs

H. D. Quednau*

Summary

A system of computer programs is presented which is able to build up design, hypothesis, and contrast matrices for any kind of experimental designs. The model on which the respective analysis is based, is described to the system by some equations whose syntax closely resembles mathematic formula language; these equations are used to build up automatically the desired matrices. The programs are written in the formula processing computer language PL/I-FORMAC and are fully compatible with PL/I-programs; an invocation from a FORTRAN-program is also possible. The practical application of the program system is demonstrated by several examples, which are based on different models in which design matrix dependent distributions are involved.

Zusammenfassung

Es wird ein System von Computerprogrammen vorgestellt, mit dessen Hilfe sich Struktur-, Hypothesen- und Kontrastmatrizen für beliebige Versuchsanlagen generieren lassen. Das Modell, das der jeweiligen Analyse zugrunde liegt, wird mit einigen Gleichungen beschrieben, deren Syntax eng an die gewohnte mathematische Formelsprache angelehnt ist; aus diesen Gleichungen werden die gewünschten Matrizen automatisch aufgebaut. Die Programme sind in der formelverarbeitenden Computersprache PL/I-FORMAC geschrieben und mit PL/I-Programmen voll kompatibel; auch ein Aufruf von FORTRAN-Programmen ist möglich. Die praktische Anwendung des Programmsystems wird an mehreren Beispielen erläutert, denen verschiedene Anwendungsfälle von Verteilungen mit Strukturmatrizen zugrunde liegen.

1. Introduction

Design matrices play an important part in the field of modern statistics. Any analysis of variance, covariance, or regression can be expressed in terms of the linear model, where the expectation of the random vector or matrix is the product of a design matrix and the vector (univariate case) or matrix (multivariate case) of design parameters. Further applications of design matrices are the general linear model, where multiplicative relations between parameters are admitted, and the analysis of Poisson or binomial variables, where the respective parameters λ_i or p_i are obtained by transforming the scalar product of line i of the design matrix and the vector of design parameters, e.g. by a logit or an arc-sine transformation. If the design parameters are fixed, then linear hypotheses can be formulated by means of the hypothesis matrix H and a result matrix R by

$$H_0 : H \cdot P = R.$$

A set L of linear contrasts between the lines of $H \cdot P$ can be formulated using contrast matrices

$$L = C \cdot (H \cdot P).$$

A linear subhypothesis concerning a subset of the lines of $H \cdot P$ can be formulated by: $H_{0s} : C_s(HP) = R_s$, where the lines of C_s are a set of independent lines of the contrast matrix C.

In the univariate case, the matrices P, R, and L reduce to the respective vectors $\vec{p}, \vec{r}$, and $\vec{l}$.

In non-standard applications, for instance if an unbalanced diallel analysis with complicated reparametrization conditions is concerned, the construction of design, hypothesis, and contrast matrices or vectors can be tedious and error prone. WILKINSON and ROGERS (1973) have published an Anova program which automatically generates the design matrix in the case of certain hierarchical and factorial designs. The present paper presents a set of computer programs which build up design, hypothesis, and contrast matrices by symbolic manipulation of some mathematical formulas that describe the model on which the analysis is based, without any restrictions concerning the kind of this model.

The computations which were necessary to develop the program system and to test out the given examples were performed on the IBM 370/168 of the computing centre at the university of Bonn.

2. The FORMAC Language

The computer programs which set up the design, hypothesis, and contrast matrices by means of formula manipulation, are written in the computer language PL/I-FORMAC (FORmula MAnipulation Compiler). This language is an extension of PL/I, such that any valid PL/I-statement remains valid also in PL/I-FORMAC. To give an idea of the structure of the language, let us consider the following FORMAC-statements:

```
LET(A(1) = X * Y + Z; A(2) = 5 * Y; R = A(1) + A(2);
C = COEFF(R; Y)); PRINT_OUT(C);
```

When these statements are executed, the character string »C = 5 + X« will be printed.

An important feature of PL/I-FORMAC is the use of CHAIN-variables, whose value is a list of atomic FORMAC-symbols. CHAIN-variables can be concatenated with one an-

* Institute of Genetics, University of Bonn, Federal Republic of Germany

other or with any other variables. Thus, the statement LET (C1 = (A, B); C2 = (C, D, E); C3 = (C1, C2, F, G, H (1), 3)) results in C3 ←′ (A, B, C, D, E, F, G, H (1), 3)′. The i^{th} element of a chain c can be referred to by the expression ARG(i, c). Thus, »LET (X = ARG (8, C3))« ⇒ X ←′ H (1)′.

PL/I-FORMAC has been defined in IBM-CORP (1967) and was first designed to cooperate with object modules produced by the PL/I-F-compiler. An enlarged version, named FORMAC 73 and published by BAHR (1977), can be used together with the products of the PL/I-OPTIMIZING or PL/I-CHECKER compilers. With the exception of the additional features in FORMAC 73, the two versions are compatible on the source level. Furthermore, we have written a set of PL/I-programs which can be used instead of the FORMAC-subroutines. These programs invoke a FORTRAN-written LISP-interpreter (a slightly modified version of the LISP1-system published by NORDSTRÖM et al. in 1973), and performs formula manipulation by means of the LISP-based system LIBAFORM (QUEDNAU 1976a, 1979). The combination LISP1-LIBAFORM needs only a FORTRAN-compiler as system dependent software, but it requires a greater amount of computer memory and more CPU-time than the FORMAC-system.

The formula manipulating abilities of PL/I-FORMAC have been used in automatic program generation for maximum likelihood estimation (QUEDNAU 1973, 1975, 1976b, c, KLEIN and QUEDNAU 1976).

3. Notational Remarks

Throughout this paper, we will write real numbers and real valued functions in small letters, vectors in small letters with an arrow above them, and matrices in capital letters. Components of vectors and matrices are characterised in the following manner:

a_i:	i^{th} component of vector $\vec{a}$;
$b_{i,j}$:	component of matrix B in line i and row j
$\vec{b}_{i,*}$:	vector formed by i^{th} line of matrix B
$\vec{b}_{*,j}$:	vector formed by the j^{th} row of matrix B.
l.$\vec{a}$, l.B.:	number of these lines of vector $\vec{a}$, or matrix B, resp
r.B.:	number of rows of matrix B

If commonly used symbols are concerned, these rules may be violated, e.g. the symbol $F_{1-\alpha}(f_1, f_2)$, which means the quantile of a F-distribution, is a scalar value. Variables of computer programs are always written with Latin capital letters.

4. Definition of a Design Matrix

Let Y be a (n x m)-random matrix, and let $\vec{t}$, called the *design index function,* assign a real vector to each line index i of Y(i = 1,.., n = 1.Y), such that $\vec{t}(i) = (t_1(i);..,t_\zeta(i))$. The components of $\vec{t}(i)$ are called the *design indices* of the i^{th} line of Y. The *design function* $\vec{d}$ assigns a real vector to the design index vector such that $\vec{d}(\vec{t}(i)) = \vec{d}(t_1(i);..,t_\zeta(i)) = (d_{1,i}i;..,d_{r,i})$. The matrix whose lines are identical to the function values of $\vec{d}$, is called the *design matrix* S. So we have:

$$\vec{s}_{i,*} = \vec{d}(\vec{t}(i)).$$

The spectral or density function of Y is supposed to depend on the matrix product of the design matrix S and a matrix P, whose components are called *design parameters*. In the most common applications, we have E(Y) = S · P or E(Y) = F(S · P), where $f_{i,j}$ may be, for instance, a logit or probit transformation. The design parameters may be fixed or random, the coefficients of the design matrix may be known or known functions of the design parameters (see chapter 6.3).

The notations introduced above may be illustrated by the model which underlies a multivariate two way factorial design without interactions. The number of levels may be a for factor (1), and b factor (2). Let $\vec{y}_{i,*}$ (= i^{th} observation = i^{th} line of Y) be taken from level j of (1) and level k of (2). Then $\vec{y}_{i,*} = \vec{y}_{j,k,*} \sim N_m(\vec{\mu}_{j,k}, \sum) \equiv N_m(\vec{\mu} + \vec{\alpha}_j + \vec{\beta}_k, \sum)$.

To make the model unambigous, the parameters are submitted to linear constraints:

$$\sum_{j=1}^{a} \vec{\alpha}_j = \sum_{k=1}^{b} \vec{\beta}_k = \vec{0} \Rightarrow \vec{\alpha}_a = -\sum_{j=1}^{a-1} \vec{\alpha}_{j\wedge}\, \vec{\beta}_b = -\sum_{k=1}^{b-1} \vec{\beta}_k.$$

The parameter set of this model comprises the matrix $P = (\vec{\mu}, \vec{\alpha}_1, \ldots, \vec{\alpha}_{a-1}, \vec{\beta}_1, \ldots, \beta_{b-1})^T$, and the vector $\vec{q}$ which consists of the (m(m+1)/2) coefficients of the upper triangular part (including diagonal) of the symmetric matrix $\sum$. The components of P are the design parameters of this model.

The two-dimensional design index function $\vec{t}$ assigns the respective level of (1) and (2) to the line indices: $\vec{t}(i) = (j,k)$. The design index vector (j,k) is transformed by the design function $\vec{d}$ into a (a + b − 1)-dimensional vector, such that $\vec{d}(j, k) \cdot P = \vec{\mu}_{j,k}$. So we get: $d_1 = 1$ for all (j, k); furthermore, if $j \neq a$ and $k \neq b$, then $d_{j+1} = d_{a+k} = 1$ and $d_c = 0$ for $c \notin \{1, j+1, a+k\}$; if $j = a$ then $d_2 = \ldots = d_a = -1$ (resulting from first constraint), and if $k = b$ then $d_{a+1} = \ldots = d_{a+b-1} = -1$ (resulting from second constraint).

For the density function φ of Y we get:

$$\varphi = \prod_{i=1}^{n} n_m(\vec{y}_{i,*}, \vec{s}_{i,*} \cdot P, \sum),$$

where n_m is the density function of the m-dimensional normal distribution.

In the case of the analysis of variance, both the design and the design index function are integer valued. If covariate data are included into the analysis, then any real numbers can figure as design indices and as components of the design matrix. Let us extend the Anova example described above by assuming, that a scalar covariate $c_{j,k}$ has been observed together with every random vector $\vec{y}_{i,*} = \vec{y}_{j,k,*}$, and that $\vec{y}_{j,k,*} \sim N_m(\vec{\mu} + \vec{\alpha}_j + \vec{\beta}_k + c_{j,k}\vec{\gamma}_1 + c_{j,k}^2\vec{\gamma}_2, \sum)$.

Table 1. Design indices, design matrix and design parameters for a hypothetical data set $(Y, \vec{c})$ in the case of a two-variate bifactorial quadratic analysis of covariance

I	Y		$\vec{c}$	t_1	t_2	t_3	design matrix								P	
1	5	8	2	1	1	2	1	1	0	1	0	0	2	4	μ_1	μ_2
2	6	9	3	2	1	3	1	0	1	1	0	0	3	9	$\alpha_{1,1}$	$\alpha_{1,2}$
3	5	9	3	3	1	3	1	−1	−1	1	0	0	3	9	$\alpha_{2,1}$	$\alpha_{2,2}$
4	7	8	3	1	2	3	1	1	0	0	1	0	3	9	$\beta_{1,1}$	$\beta_{1,2}$
5	8	7	4	2	2	4	1	0	1	0	1	0	4	16	$\beta_{2,1}$	$\beta_{2,2}$
6	6	10	4	3	2	4	1	−1	-1	0	1	0	4	16	$\beta_{3,1}$	$\beta_{3,2}$
7	8	12	4	1	3	4	1	1	0	0	0	1	4	16	$\gamma_{1,1}$	$\gamma_{1,2}$
8	4	6	5	3	3	5	1	−1	−1	0	0	1	5	25	$\gamma_{2,1}$	$\gamma_{2,2}$
9	7	9	5	1	4	5	1	1	0	−1	−1	−1	5	25		
10	8	10	5	2	4	5	1	0	1	−1	−1	−1	5	25		
11	5	8	5	3	4	5	1	−1	−1	−1	−1	−1	5	25		

Now, $\vec{t}(i) = (j, k, c_{j,k})$. As for the design function $\vec{d}$, we get for d_1 through d_{a+b-1} the same values as in the Anova model, furthermore $d(j, k, c_{j,k})_{a+b} = c_{j,k}$, and $d(j, k, c_{j,k})_{a+b+1} = c_j^2{}_{,k}$.

Table 1 shows the design indices, the design matrix, and the design parameters for this extended model and a hypothetical data set $(Y, \vec{c})$.

5. The Formula Manipulating Programs

5.1 Definition of a simple model

The usage of the formula manipulating programs will best be described by an example, which is not too complicated, but which, on the other hand, cannot be handled by special purpose programs. The model which will be used here was proposed by TOPHAM (1966), making use of former work by HAYMAN (1954) and WEARDEN (1964). Some variations of this model are described by WALTERS and GALE (1977). TOPHAM writes the expectation of a metric trait in a cross between a female of genotype j and a male of genotype k by

$$E(y_i = y_{j,k,r}) = \mu + g_j + g_k + m_j + w_{j,k} + n_{j,k}$$

where r is the replication index, μ represents the grand mean, g_j and g_k are the parental effects, $w_{j,k}$ the interactions between them, m_j the maternal effects and $n_{j,k}$ the interaction between maternal and parental effects. The effects may be fixed or random, the observation y may be a scalar or a vektor. The parameters are submitted to the following constraints:

$w_{j,k} = w_{k,j}$, $n_{j,k} = -n_{k,j}$, $n_{j,j} = 0$ for all (j, k).

If the effects are fixed, we have moreover

$$\sum_l g_l = \sum_l m_l = \sum_j w_{j,k} = \sum_k w_{j,k} = \sum_j n_{j,k} = \sum_k n_{j,k} = 0.$$

In the example model of this chapter, we will extend TOPHAM's model by assuming, that $E(y_i)$ depends not only on the genetic parameters, but moreover on a covariate $x_i = x_{j,k,r}$, so that we finally have

$$E(y_i) = E(y_{j,k,r}) = \mu + g_j + g_k + m_j + w_{j,k} + n_{j,k} + b \cdot x_i$$

5.2 The subprogram GETFMC

The routine GETFMC, which reads the information about the design to be analysed from an input file, is invoked by the programs GETMODL and GETHYP. The input to GETFMC is a series of statements, each of which either corresponds to the rules of a FORMAC-assignment within a LET-command, or is a DO-group of the form »DO control variable = lower bound TO upper bound; st_1; ...; st_n; END;«. st_1 through st_n may again be either FORMAC-compatible assignments or DO-groups. When a statement has been read, its information is passed to the FORMAC-system by invoking »CALL DENFMC1 (›statement‹)« once or, in the case of a DO-loop, several times. When the DO-loop has been finished, the control variable is reset to its symbolic value, i.e. it is atomized. Reading a %-sign causes the program to return. Thus, if we have on the input file

```
L = 10; K = 5; DO I = 1 TO L; A(I, K) = 0; DO J = 1 TO K
- 1;
A(I, K) = A(I, K) - A(I, J); END; END; %
```

then, after return from GETFMC, we get

A(1, 5) ← − A(1, 1) − A(1, 2) − A(1, 3) − A(1, 4), ..., A(10, 5) ← − A(10, 1) − A(10, 2) − A(10, 3) − A(10,4).

5.3 The subprogram GETMODL

The Programm GETMODL first reads in the *design generation formula* d g f, which describes the vector products $\vec{s}^T_{i,*} \cdot P$ in symbolic form, using the symbols of the design parameters and the design indices. The design generation formula is stored in an external character string variable, then the subroutine GETFMC is invoked. After return from GETFMC, all necessary information about the model, the data structure and possible program switches are supposed to be known to the system. The following assignments to FORMAC-variables are required:

1 The CHAIN-variable P≠ LIST must contain the symbols of the design parameters.
2 The symbols of the design indices must be elements of the CHAIN-variables S≠ LIST or C≠ LIST.
3 The FORMAC-variable NTOT contains a number, which is equal to the lines of Y; if Y contains more than one row, then the number of rows must be stored in NMAN. Other assignments are optional and will be described when required.

If a set of data is to be analysed following the model of TOPHAM's with fixed effects extended by assuming linear dependence on a covariate, then the user must supply the input given in table 2.

The program GETMODL passes information about the dimensions of Y and P and the number of design indices to the invoking program by integer valued, PL/I-compatible arguments.

5.4 The subprogram GENDES

The program GENDES evaluates the design matrix in a two-dimensional array, which is passed by the invoking program as an argument. An option code argument indicates, if the data is already in computer memory (in two 2-dimensional argument arrays, one for the observation matrix Y and one for the design indices), or if they have to be read in. In the latter case, the data items have to be written on the input file in the following manner: First stand the current values for those design indices which are listed in S≠ LIST, preceded by a repetition number n which declares, how many observations belong to this design index combination. Now are following these n observations, each of them followed by the current values of those design indices which are listed in C≠ LIST. Suppose we have got 3 observations from the cross of genotype 5 (mother) with genotype 7 (father), each observation consisting of a bivariate random

Table 2. Input to program GETMODL, defining TOPHAM's model

```
MY + G(J) + G(K) + M(J) + W(J,K) + N(J,K) + B * X;
NN = < number of genotypes involved in the diallel analysis >;
DO J = 2 TO NN ; DO K = 1 TO J - 1 ; W(J,K) = W(K,J) ;
N(J,K) = - N(K,J) ; END ; END ;
G(NN) = 0 ; M(NN) = 0 ; DO L = 1 TO NN - 1 ; G(NN) = G(NN) - G(L) ;
M(NN) = M(NN) - M(L) ; END ;
DO K = 1 TO NN-1 ; W(NN,K) = 0 ; N(NN,K) = 0 ; DO J = 1 TO NN - 1 ;
W(NN,K) = W(NN,K) - W(J,K) ; N(NN,K) = N(NN,K) - N(J,K) ; END ; END ;
DO J = 1 TO NN; W(J,NN) = 0 ; N(J,NN) = 0 ; DO K = 1 TO NN - 1 ;
W(J,NN) = W(J,NN) - W(J,K) ; N(J,NN) = N(J,NN) - N(J,K) ; END ; END ;
S≠ LIST = (I,J) ; C≠ LIST = (X) ; P≠ LIST = (B,MY) ;
DO L = 1 TO NN - 1 ; P≠ LIST = (P≠ LIST,G(L)) ; END ;
DO L = 1 TO NN - 1 ; P≠ LIST = (P≠ LIST,M(L)) ; END ;
DO J = 1 TO NN - 1 ; DO K = J TO NN - 1 ;
P≠ LIST = (P≠ LIST,W(J, K)) ; END ; END ;
DO J - 1 TO NN - 2 ; DO K = J + 1 TO NN - 1 ;
P≠ LIST = (P≠ LIST,N(J,K)) ; END ; END ;
NTOT = < number of univariate or multivariate observations >;
NMAN = < number of rows of Y, required only in case of a multivariate analysis >; %
```

variable and a covariate. The input stream describing a hypothetical data item could then be 3 5 7 19.2 6.2 200 20.5 7.3 250 25.4 8.2 300, where (19.2, 20.5, 25.4) refer to the first variate, (6.2, 7.3, 8.2) to the second variate, and (200, 250, 300) to the covariate. If the repetition number is equal for all S# list design combinations, it need occur only once at the beginning of the input file. Its overall validity is then indicated by a negative sign. – If, for some or all of the S# LIST design indices the current values can be evaluated by a DO-loop, they need not be explicitly listed in the input. In this case, the upper bounds of the indices must be listed in the CHAIN-variable K# LIST. If the corresponding lower bounds are not all equal to 1, they must be contained in the CHAIN-variable L# LIST. Suppose the diallel analysis comprises 10 genotypes, such that the indices j and k range from 1 to 10. This can be indicated by writing »K# LIST = (10,10);« in the input for GETMODL, and in consequence the design indices J and K will be automatically assigned values from 1 to 10 by a nested DO-loop, J being the control variable of the outer loop, K being changed in the inner loop. If for some combination of J and K, observation data are completely missing, then the corresponding data item consists only of the repetition number 0. – If the FORMAC-variable Y# TRANS is assigned a value of the form 'f(#Y#)', then the random variates will be transformed according to the transformation expressed by f, before they are stored. Thus, the assignment »Y# TRANS = LOG(#Y#/(1 − #Y#)« causes a logit transformation to be performed on each variate.

When the design-indices are set to their current values, either automatically by DO-loops or by assigning to them the corresponding input data, then the elements of the current line of the design matrix are calculated. This is done by evaluating the coefficients of each design parameter within the design generation formula. When data from a cross between genotypes 5 (father) and 7 (mother) are concerned, and the corresponding covariate has the value x = 200, then the design generation formula gets the form

»MY + G(5) + G(7) + M(5) + W(5, 7) + N(5, 7) + B · 200«.

Thus, the rows corresponding to parameters μ, g_5, g_7, m_5, $w_{5,7}$, and $n_{5,7}$ will be set to 1, the row corresponding to parameter b will be set to 200, all other rows will contain zeros.

5.5 *The subprogram GETHYP*

The subprogram GETHYP, by invoking GETFMC, reads in the information which is necessary to build up the hypothesis, result, contrast, and subhypothesis matrices. After return from GETHYP, the following assignments are assumed to have been performed.

1 The CHAIN-variable H# LIST has been assigned a list of linear combinations of the design parameters, such that the elements of H# LIST correspond to the lines of the hypothesis matrix H which is to be generated. If the result matrix R is not equal to zero, then either the CHAIN-variable G# LIST contains the elements of the result vector $\vec{r}$ (univariate case), or the indexed CHAIN-variable R# LIST (i) contains the i^{th} row of the result matrix R (general case).

2 If contrast matrices shall be generated, then the CHAIN-variable @# LIST and/or the indexed CHAIN-variable ## LIST(k) have been assigned a list of linear combinations of the design parameters, which must be linear dependent on the lines of H · P. Alternatively, the i^{th} line of H · P can be referred to as »###(i)«. In general, the invoking program will use the contrast matrix defined by @# LIST to construct confidence intervals for the respective contrasts, and by the contrast matrices defined by ## LIST(k), it will formulate subhypotheses of the form: $H_{os} : C_s \cdot (H \cdot P) = R_s$. The value for the result-matrices can be stored in the CHAIN-variables G## LIST(k) (univariate case), or R## LIST(k,i) (general case) in the same way as it is described for the result matrix R and the CHAIN-variables G# LIST and R# LIST(i).

Suppose we want to test, in our example model, if any maternal effects, including interactions between maternal and parental effects, are present. Then, the input to GETHYP would be: H# LIST = M(1); DO I = 2 TO NN-1; H# LIST = (H# LIST, M(J)); END;
DO J = 1 TO NN-2; DO K = J + 1 TO NN-1; H# LIST = (H# LIST, N(J, K)), END; END; %

If we want to test that $\vec{g}_1 = \vec{g}_2 + (5, 2) \wedge \vec{g}_5 = \vec{g}_7$,

then we write: H# LIST = (G(1) − G(2), G(5) − G(7); R# LIST(1) = (5, 0); R# LIST(2) = (2, 0); %.

The following example demonstrates, how a set of linear contrasts can be generated, which will be used to construct a confidence band for the response of cross (5 × 7), the independent variable ranging from 100 to 300. In this case, an univariate model is supposed.

H# LIST = (G(5) + G(7) + W(5, 7) + M(5) + N(5, 7), B);
@# LIST = (###(1) + 100 ∗B); DO L = 1 TO 20;
@# LIST = (@# LIST, ###(1) + (100 + 10 ∗L)∗ B); END;
%

An example for the usage of ## LIST(k) will be given in chapter 6.1.

GETHYP passes the number of elements of the CHAIN-variables H# LIST, R# LIST (or G# LIST, resp.) and @# LIST as well as the range of k in ## LIST(k) to the invoking program, such that enough space can be provided for the corresponding hypothesis, result, and contrast matrices.

5.6 *The subprogram GENHYP*

The program GENHYP builds up the hypothesis, result, and contrast matrices. The j^{th} element of the i^{th} line of H is generated by performing : COEFF (ARG(I, H# LIST), p_j). If contrast matrices are to be constructed, the program first builds up a provisional matrix C* by : $c^*_{i,j}$ = COEFF(ARG(I, list), p_j), where list may be @# LIST or ## LIST(l), such that C* · P = C · (H · P). If the contrasts are formed correctly, they can be expressed as linear combinations of the lines of H · P, and we have : $C = C^* \cdot H^T(HH^T)^{-1}$.

For the construction of the result matrix, only the elements of R# LIST or G# LIST have to be changed from FORMAC numbers into PL/I numeric variables.

6. Classes of Statistical Models in which Design Matrices are Involved

In this chapter, three specific cases of design matrix dependent distributions will be presented. Each of them is exemplified by a model of biological importance. Moreover, it will be described which kinds of biometric analyses can be performed using the respective design and related matrices. Only such algorithms will be referred to, which are either published as computer programs, or for which programs can easily be made.

6.1 The general linear model with fixed effects

If the lines of Y are independent and distributed like: $\vec{y}_{i,*} \sim N_m(\vec{s}_{i,*} \cdot P)^T, \sum)$, where the covariance matrix $\sum$ is equal for all line indices i, the design matrix S is completely known, and P contains only fixed design parameters, then we have a realization of the general linear model with fixed effects. In this case, efficient estimates for the parameters are obtained by

$\vec{p}_{*,j} : \vec{r}^T_{*,j} \cdot \vec{r}_{*,j} = (S \cdot \vec{p}_{*,j} - \vec{y}_{*,j})^T (S \cdot \vec{p}_{*,j} - \vec{y}_{*,j}) \stackrel{!}{=} \text{Min};$

$\sum = (n-k)^{-1} r^T R,$

where k = l.P = number of design parameters. If S is well conditioned, $\hat{P}$ can be calculated by: $\hat{P} = (S^TS)^{-1} S^TY$, otherwise an orthogonalization method can be used (GEIDEL and PRECHT 1978). For the m columns of the residual matrix R we have: $\vec{r}_{*,j} \sim N_n(\vec{0}, \sigma_{j,j} \cdot Q)$, where $Q = I_n - S(S^TS)^{-1} S^T$, j = 1, ..., m, and $\sigma_{j,j}$ is the j^{th} element of the diagonal of $\sum$.

With the help of Y, S, and Q, a test can be set up to prove the validity of the supposed model: If the model is correct, then any orthonormal base B of Q will transform the vectors of residuals into random vectors, whose components are independent, and have expectation 0 and equal variance:

$B \cdot \vec{r}_{*,j} \sim N_{n-m}(\vec{0}, \sigma^2_c \cdot I_{n-m})$

So, the validity of the model can be tested by proving the normality of $B \cdot \vec{r}_{*,j}$. This can be done by the well-known χ^2-test or by methods which make use of the distribution of the 3rd and 4th sample moments of a normally distributed variable (GEBHARDT 1966, d'AGOSTINO and PEARSON 1973, and BOWMAN 1973). An orthonormal base can be obtained by $B = D^{-1/2} M^T$, where M is composed by the (n-m) eigenvectors of Q, and the diagonal matrix D contains the respective eigenvalues. PUTTER (1967) describes the construction of orthonormal bases which are suitable for certain special designs.

Whereas the method of orthonormal bases allows for testing the global validity of the model, other procedures have been described to detect single outliers within the data set. ANSCOMBE and TUKEY (1963), ANDREWS (1971), and BEHNKEN and DRAPER (1972) suggest using the studentized residuals $t_{i,j} = r_{i,j} / \sqrt{\sigma_{j,j} q_{i,i}}$, where $\sigma_{j,j}$ and $q_{i,i}$ mean the j_{th} diagonal element of $\sum$ and the ith diagonal element of Q, resp. LUND (1975) worked out an easily practicable test, which is based on papers by ELLENBERG (1973) and PRESCOTT (1975): If no outlier exists in $\vec{y}_{*,j}$, then, with probability $(1-\alpha)$, the absolutely largest t_i is less than a tabulated value which depends on α, n, and m. COOK (1977) presents a function of S, Y and i, whose value measures, for each line number i, how far a hypothetical outlier $y_{i,j}$ would influence the estimate of $\vec{p}_{*,j}$.

A linear hypothesis concerning the design parameter matrix P can be formulated by means of the hypothesis matrix H and the result matrix R : H_o : HP = R. A test of H_o is based on a test function of the random variables $SP_{rest} = (n-k) \cdot \sum$ and $SP_{hyp} = (HP - R)^T [H(S^TS)^{-1} H]^{-1} (HP - R)$.

If S is so ill conditioned that $(S^TS)^{-1}$ cannot be obtained, then SP_{hyp} can be calculated by: $SP_{hyp} = SP_{rest} - SP_o$, where SP_o is the SP_{rest} of the model restricted by H_o. Generally, SP_o will be obtained in a second run of the same program, which has calculated the design parameters and the SP_{rest} for the full model.

For the multivariate case, several different test functions have been proposed. KRES (1975) gives a survey on the multivariate methods developed so far, WEILING and UNGER (1977) have compared their respective efficiency. For the variances $\sigma_{j,j}$, simultaneous confidence intervals can be constructed using the method of JENSEN and JONES (1969).

In the univariate case, the random variables $\sigma^{-2} SP_{hyp}$ and $\sigma^{-2} SP_{rest}$ are independent and have a χ^2-distribution with z and (n − k) degrees of freedom, where $\sigma^2 = \sigma_{1,1}$ and z is the number of lines of H. From this follows that

$\frac{SP_{hyp}}{SP_{rest}} \cdot \frac{n-k}{z} \sim F(z, n-k)$, and, with probability $(1-\alpha)$, the following inequation holds:

$(\vec{g} - \hat{\vec{g}}^T Q (\vec{g} - \hat{\vec{g}}) - \zeta \le 0$, where $\vec{g} = H\vec{p}, \hat{\vec{g}} = H \cdot \hat{\vec{p}},$

$Q = [H(S^TS)^{-1}H^T]^{-1}$, and $\zeta = F_{1-\alpha}(z, n-m) \cdot z \cdot \hat{\sigma}^2$.

Using this inequality we can 1 test linear subhypotheses of H_o, where the probability of rejecting at least one true subhypothesis, is at most equal to α, and 2 construct simultaneous confidence intervals for any linear contrasts of $\vec{g}$.

Let C be a matrix with r · C = z, l.C = w < z and rank (C) = l.C, such that $C \cdot (H \cdot \vec{p}) = C \cdot \vec{g}$ is a set of independent linear contrasts of $\vec{g}$. To test the subhypothesis : $H_{os} : C \cdot \vec{g} = \vec{r}_s$, we have to find

$a_s = \min_{\tilde{\vec{g}}/Con} (\tilde{\vec{g}} - \hat{\vec{g}})^T Q (\tilde{\vec{g}} - \hat{\vec{g}})$, where the constraint Con implies: $C \cdot \tilde{\vec{g}} - \vec{r}_s = \vec{0}$.

If $a_s > \zeta$, then H_{os} is rejected, because in this case, there is no vector $\tilde{\vec{g}}$ inside the confidence interval of $\vec{g}$ which fulfils the equation imposed by H_{os}.

The constraint minimum $a_s(\tilde{\vec{g}})$ can be transformed to an unconstrained minimum with the help of Lagrangian multipliers:

$a_s = \min(l(\tilde{\vec{p}}, \vec{\lambda})) = \min_{\tilde{\vec{p}}, \vec{\lambda}} (\tilde{\vec{g}} - \hat{\vec{g}})^T Q (\tilde{\vec{g}} - \hat{\vec{g}}) +$

$\sum_{s=1}^{w} \lambda_s \cdot \left[\left(\sum_{t=1}^{z} c_{s,t} \cdot \tilde{\vec{g}}_t \right) - r_s \right]$

$\text{Since } \frac{\partial l}{\partial \tilde{g}_i} = 2 \cdot \sum_{t=1}^{z} q_{i,t} (\tilde{g}_t - \hat{g}_t) + \sum_{s=1}^{w} \lambda_s c_{s,i}$

$\text{and } \frac{\partial l}{\partial \lambda_s} = \sum_{t=1}^{z} (c_{s,t} \tilde{g}_t) - r_s,$

the minimization of l, which is identical to the determination of the root of »$\frac{dl}{d(\tilde{\vec{g}}, \vec{\lambda})} = \vec{0}$«, leads to a system of linear equations which can readily be solved:

$\begin{pmatrix} 2 \cdot Q & C^T \\ C & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{g}^* \\ \vec{\lambda}^* \end{pmatrix} = \begin{pmatrix} 2 \cdot Q \cdot \hat{\vec{g}} \\ \vec{r} \end{pmatrix}$

Simultaneous confidence intervals of any linear contrasts $\vec{c}^T \cdot \vec{g}$ can be obtained by optimizing $\vec{c}^T \cdot \tilde{\vec{g}}$ under the constraint.

$Con : (\tilde{\vec{g}} - \hat{\vec{g}})^T \cdot Q \cdot (\tilde{\vec{g}} - \hat{\vec{g}}) - \zeta = 0$. The constrained extreme values of $\vec{c}^T \tilde{\vec{g}}$ are:

$\min, \max_{\tilde{\vec{g}}/Con} \vec{c}^T \cdot \tilde{\vec{g}} = \vec{c}^T \cdot \hat{\vec{g}} \pm \sqrt{\zeta \cdot \vec{c}^T Q \vec{c}}$ (SCHEFFÉ 1953).

BOUDEN (1970) gives a survey on further sets of confidence intervals, which can be constructed using S and H in the case of the univariate general model.

If the variances of the y_i are not equal, but their ratios are known, then we can achieve homoskedasticity by a linear transform, as we have

$y_i \sim N(\vec{s}_{i,*}^T \cdot \vec{p}, w_i \cdot \sigma^2) \Rightarrow w_i^{-0,5} \cdot y_i \sim N((w_i^{-0,5} \cdot \vec{s}_{i,*})^T \vec{p}, \sigma^2).$

In a similar way, a multivariate random variable can be transformed into a set of approximately independent univariates with variances equal to unity, using $\hat{\Sigma}$ as weight matrix (WOTTAWA 1974).

When the design and hypothesis matrices have been set up, the numerical evaluation of the analysis can be performed using some standard system of computer programs. Critical surveys on algorithms and published computer programs are given by NOWAK (1975), PRECHT (1978) and GEIDEL and PRECHT (1978), moreover the system of BRYCE and CARTER (1974) should be mentioned.

If a reparametrization has been omitted such that S^TS may be singular, the analysis can be performed as described by NOLLAU (1975) with the help of RAO's (1962) generalized inverse. A simple FORTRAN program to calculate the generalized inverse has been published by RUST et al (1966). Investigations on estimable linear combinations in the case of a singular S^TS have been done by BERCHTHOLD (1977).

The distance between a hypothesis $H_{oj}: H \cdot \vec{p} = \vec{g}_{oj}$ and the respective true statement can be measured by the expression $\sqrt{n^{-1}}\delta_j$ with

$$\delta_j = (H \cdot \vec{p} - \vec{g}_{oj})^T H(S^TS)^{-1} H^T (H\vec{p} - \vec{g}_{oj})$$ (BULMER 1958).

Simultaneous confidence intervals for any number of δ_j, which are valid together with the Scheffé-intervals for linear contrasts, are obtained by:

$$\max\left[0, (\sqrt{\hat{\delta}_j} - \sqrt{c})^2\right] \leq \delta_j \leq (\sqrt{\hat{\delta}_j} + \sqrt{c})^2,$$

where $c = z \cdot \hat{\sigma}^2 \cdot F_{1-\alpha}(z, n-k)$. A shorter, nonsimultaneous interval for a special δ can be obtained using the non-central F-distribution, which can be approximated by the central F-distribution by transformation rules given by PATNAIK (1949).

If two hypotheses H_{o1} and H_{o2} shall be tested using the same set of observational data, then the independence of H_{o1} and H_{o2} can be proved in the following way: The user formulates the compound hypothesis

$H_{o1} \wedge H_{o2}: \frac{H_1}{H_2} \cdot \vec{p} = \frac{\vec{g}_1}{\vec{g}_2}$ and has the matrix

$U = \mathrm{Cov}\left(\frac{H_1}{H_2} \cdot \hat{\vec{p}}\right) = \frac{H_1}{H_2} \cdot (S^TS)^{-1} \cdot (H_1{}^T : H_2{}^T)$ printed out. If H_{o1} and H_{o2} are independent, then $u_{i,j} = 0$ for all (i, j) with $i > z_1 \wedge j \leq z_1$ or $j > z_1 \wedge i \leq z_1$, where z_1 is the line number of H_1.

FORKMAN and SEYFFERT (1977) have developed a model to describe the expectation of a quantitative trait which is dependent on several genes. In the case of 3 loci L(a), L(b), L(c), each of which have two alleles g^r and g^d (g = a, b, c) they get: $E(y_i = y_{j,k,l}) = z + a_j + b_k + c_l + (ab)_{j,k} + (ac_{j,l} + (bc)_{k,l} + (abc)_{j,k,l}$, j, k, l ε {0,1,2} (model 2 of the cited paper). The indices of the effects of the 3 gene loci are 0, if individual i is homozygous for gene g^r in the respective locus, they are 1 if it is heterozygous, and they are 2 if it is homozygous for g^d. The effects of the alleles g^r are taken as reference values, as well for the main effects as for the interactions. Therefore all parameters which have a zero in their index set, are zero by definition:

$a_o = b_o = c_o = (ab)_{o,*} = (ab)_{*,o} = (ac)_{o,*} = (ac)_{*,o} = (bc)_{o,*} = (bc)_{*,o} = (abc)_{o,*,*} - (abc)_{*,o,*,*} = (abc)_{*,*,o} = 0.$

The parameter z is thus defined as the expected value of an individual with genotype (a^r, b^r, c^r).

The model can readily be extended to more than 3 gene loci and more than 2 alleles per gene locus; moreover parameters can be added which account for environmental effects and interactions between environment and genotype.

Table 3. Statements defining FORKMAN's and SEYFFERT's model

a Input to program GETMODL

```
Z + A(J)  + B(K) + C(L) + AB(J,K) + AC(J,L) + BC(K,L)
          + ABC(J,K,L) ;
S#LIST = (J,K,L) ;
P#LIST = (Z,A(1),A(2),B(1),B(2),C(1),C(2)) ;
DO L1 = 1 TO 2 ; DO L2 = 1 TO 2 ;
P#LIST = (P#LIST,AB(L1,L2),AC (L1,L2),BC(L1,L2)) ; END ; END ;
DO J = 1 TO 2 ; DO K = 1 TO 2 ; DO L = 1 TO 2 ;
P#LIST = (P#LIST,ABC(J,K,L)) ; END ; END ; END ;
NTOT = < number of data > ; %
```

b Input to program GETHYP

```
H#LIST = (0,25*A (2) - 0.5*A(1) , 0.25*B(2) - 0.5*B(1) ,
  0.25*C(2) - 0.5*C(1)) ;
DO L1 = 1 TO 2 ; DO L2 = 1 TO 2 ;
H#LIST = (H#LIST,AB(L1,L2),AC(L1,L2),BC(L1,L2)) ; END ; END ;
DO J = 1 TO 2 ; DO K = 1 TO 2 ; DO L = 1 TO 2 ;
H#LIST = (H#LIST,ABC(J,K,L)) ; END ; END ; END ;
##LIST(1) = ( ### (1),### (2),### (3)) ;
##LIST(2) = ( ### (4)) ;
DO I = 2 TO 20 ; ##LIST(2) = ( ##LIST(2),### (3 + I)) ; %
```

Table 3a gives the input to program GETMODL, which describes the model for our program system.

According to FISHER (1918), genetic effects can be devided into 3 groups: 1) the mean values of the alleles (= additive effects), 2) interactions between alleles (= dominance effects), 3) non-allelic interactions (= epistatic effects).

For a breeding program, it is of utmost interest to have an idea of the amount of each of these effects (WATKINS and SPANGELO 1968).

In the model cited here, component (1) comprises the three linear combinations $0.25\ g_2 + 0.5\ g_1$, component (2) is represented by $g_2 - 2g_2$, and component (3) is the set of interactions {(ab), (ac), (bc), and (abc)}.

Suppose we want to test if there is any deviance from a purely additive model of inheritance, caused either by dominance or epistatic effects. If the test value turns out to be significant, we want to test the subhypotheses, if the non-additivity can be explained by the presence of dominance or epistatic effects, or both of them. The corresponding input to program GETHYP is described in table 3 b.

If the input data y_i are means of random samples of respective size n_i, such that $y_i \sim N(\vec{s}_{i,*}{}^T \cdot \vec{p}, n_i{}^{-1}\sigma^2)$, then the first input item to GETMODL must be modified to »(design generating formula of the unweighted model) *SQRT(N)«, and the following statements are to be included: »Y#TRANS = #Y# *SQRT(N); S#LIST = (S#LIST, N);« The sample size n is then supposed to precede the corresponding observation data as last design index.

6.2 The univariate general linear model with mixed effects

Let the random vector $\vec{y}$ be distributed like:

$$\vec{y} \sim N_n\left(D\vec{\alpha} + \sum_{j=1}^{q} U_j\vec{b}_j + \sum_{k=1}^{r} W_k\vec{c}_k, \sigma^2 I_n\right)$$

The matrices D, U_j, and W_k are known, $\vec{\alpha}$ is a vector of fixed parameters, $\vec{b}_j$ and $\vec{c}_k$ are mutually independent random vectors where

$\vec{b}_j \sim N(\vec{0}, \lambda_j \cdot \sigma^2 I_{l.\vec{b}_j})$ and $\vec{c}_k \sim N(\vec{0}, M)$.

Now we can write the distribution of $\vec{y}$ as

$$\vec{y} \sim N_n(D\vec{\alpha}, \sigma^2(I_n + \sum_{j=1}^{q} \lambda_j U_j^T U_j) + \sum_{k=1}^{r} W_k M W_k^T).$$

Thus, matrix $S = (D : U_1 : \ldots : U_q : W_1 : \ldots : W_r$ is a design matrix as defined in chapter 4.

For the design parameters we get : $\vec{p} = (\vec{\alpha}, \vec{b}_1, \ldots, \vec{b}_q, \vec{c}_1, \ldots, \vec{c}_r)$, and the set of true parameters is $\{\vec{\alpha}, \sigma^2, \vec{\lambda}, \{m_{i,j}\}_{j \le i}\}$.

A simultaneous unbiased estimation of all parameters of this model ist not yet possible. RAO (1972) has published a method for an unbiased estimation of $\vec{\lambda}$ and M, which he called MINQUE (MInimum Norm Quadratic Unbiased Estimation). He suggests to estimate $\vec{\alpha}$ by weighted least squares, where the weight matrix is the inverse of

$$G = \hat{\sigma}^2(I_n + \sum_{j=1}^{q} \tilde{\lambda}_j U_j^T U_j) + \sum_{k=1}^{r} W_k \hat{M} W_k^T$$

such that $\hat{\vec{\alpha}} = (\vec{y}^T G^{-1} \vec{y})^{-1} (\vec{y}^T G^{-1}) \vec{y}$.

When $M = 0$, then point and interval estimations for the remaining parameters can be performed by the method of maximum likelihood, as described by HARTLEY and RAO (1967). The corresponding algorithms can be considerably simplified using the »W-transform«, which was introduced by HEMMERLE and HARTLEY (1973) and improved by THOMPSON (1975) so far that computer programs can readily be designed which do not require excessive CPU-time and memory space. Alternative methods of estimating the parameters by using the design matrix are found in papers by CORBEIL and SEARLE (1976a, b), which are based on a transform suggested by PATTERSON and THOMPSON (1971), by HENDERSON (1953), whose methods were extended by SEARLE (1968), and by CUNNINGHAM and HENDERSON (1968). For Henderson's »method 3« a FORTRAN program is available, which has been presented by NOLLAU (1976). DEMPFLE et al. (1977) have investigated the relative efficiency of these methods, using data from an animal breeding programme.

URFER and THÖNI (1976) applied RAO's MINQUE to estimate the parameters of growth curves. Extending their model by underlying a two-factorial design with one growth curve measured for each factor combination, and assuming linear-quadratic dependence on the covariate, we get for the distribution of $\vec{y}$:

$$y_i = y_{j,k,r} \sim N(\mu + \alpha_j + b_k + c_{o,j,k} + c_{1,j,k} \cdot x_{j,k,r} + c_{2,j,k} \cdot x_{j,k,r}^2, \sigma^2),$$

where the $x_{j,k,r}$ are known covariates, μ and α_j are fixed, $\vec{b}$ and C are mutually independent random variables with $\vec{b} \sim N(\vec{0}, \lambda \cdot \sigma^2 \cdot I_{l.\vec{b}})$ and $c_{*,j,k} \sim N(\vec{0}, M)$. This model is described to the program GETMODL by the input given in table. 4.

Table 4. Input to program GETMODL, defining a mixed effect model

```
MY + ALFA(J) + B(K) + C(0,J,K) + C(1,J,K)*X + C(2,J,K)*X**2 ;
S≠LIST = (J,K) ; C≠LIST = (X) ;
JJ = < upper bound of j > ; KK = < upper bound of k > ;
K≠LIST = (JJ,KK) ;
ALFA(JJ) = 0 ; P≠LIST = (MY) ;
DO J = 1 TO JJ - 1 ; ALFA(JJ) = ALFA(JJ) - ALFA(J) ;
        P≠LIST = (P≠LIST,ALFA(J)) ; END ;
DO K = 1 TO KK ; P≠LIST = (P≠LIST,B(K)) ; END ;
DO J = 1 TO JJ ; DO K = 1 TO KK ;
P≠LIST = (P≠LIST,C(0,J,K),C(1,J,K),C(2,J,K)) ; END ; END ;
NFIX = JJ ; NQ = 1 ; NJ(1) = KK ; NR = JJ * KK ; NK = 3 ; %
```

The last assignments pass information about the structure of the design matrix to the FORMAC-system, which can be used by the estimation program.

The meaning of the respective symbols is as follows: NFIX = number of rows of D, NQ = number of matrices U_j, NJ(j) = dimension of $\vec{b}_j$, NR = number of matrices W_k, NK = dimension of vectors $\vec{c}_k$ = l.M = r.M.

6.3 The generalized linear model

Based on investigations on SCHEFFÉ's (1959), MILLIKAN and GRAYBILL (1970) developed a generalization of the univariate linear model which admits multiplicative combination of design parameters. Suppose we have:

$$\vec{y} \sim N_n(T \cdot \vec{b} + U \cdot \vec{a}, \sigma^2 I_n).$$

The matrix T is known, the coefficients of matrix U are known functions of $\vec{b}$, moreover let rg (T^TT) = r.T and rg (U^TU) = r.U. So we get a design matrix $S = (T : U)$ which corresponds to the vector of the design parameters $\vec{p} = (\vec{b}^T : \vec{a}^T)$. The null hypothesis $H_o : \vec{a} = \vec{o}$ can be tested in the following way: Let $\hat{\vec{b}}^r$ be the restricted estimate of $\vec{b}$, which results if $\vec{a} = \vec{o}$ is assumed, i.e. $\hat{\vec{b}}^r = (T^TT)^{-1} T^T \vec{y}_r$, and let $\hat{U}$ be the matrix which is obtained by replacing $\vec{b}$ by $\hat{\vec{b}}^r$ in U.

If the test for $\vec{a} = \vec{0}$ is now performed in the way which was described for the general liner model (see chapter 6.1), but using $(T : \hat{U})$ instead of $(T : U)$ as design matrix, then we get for the test variable

$$\frac{SP_{hyp}}{SP_{rest}} \cdot \frac{n - l.\vec{b} - l.\vec{a}}{l.\vec{a}} \sim F(l.\vec{a}, n - l.\vec{b} - l.\vec{a}),$$

if H_o is true. If $\vec{a} \neq 0$, the distribution of the test variable is in general not known.

Several authors use the formal estimates $\hat{\vec{a}}^f$, which are obtained by using $(T ; \hat{U})$ as design matrix, as true estimates for $\vec{a}$. If $\vec{a} \neq \vec{0}$, then $E(\hat{\vec{a}}^f) \neq \vec{a}$, however (SPRENT 1969), although the bias is tolerable when the design is not too ill conditioned (HARDWICK and WOOD 1972).

A model with multiplicative relations between parameters was already used by YATES and COCHRAN (1938) to investigate the linear dependence of interactions on main effects. TUKEY (1949) worked out a test where the interactions in a two way factorial design are presupposed to be proportional to the main effects:

$$E(y_i = y_{j,k}) = \mu + \alpha_j + \beta_k + G \cdot \alpha_j \beta_k.$$

The model designed by MANDEL (1961) assumes that, for each level of factor (1), the interaction is linear dependent on the main effect of factor (2), i.e.

$$E(y_i = y_{j,k}) = \mu + \alpha_j + \beta_k + \theta_j \beta_k,$$
$$\text{where } \sum_j \theta_j = 0.$$

For matrix U and vector $\vec{a}$ we get for

Model TUKEY: $\vec{a} = (G)$, $u_{i,1} = \alpha_j \beta_k$

Model MANDEL: $\vec{a} = \vec{\theta}$, $u_{i,v} = \beta_k$ if v = j, o elsewhere. j and k are the design indices of i: $(j, k) = \vec{t}(i)$.

These models are the best known applications of the generalized linear model, especially MANDEL's model is often used to describe the interactions of genotypic and environmental effects.

Table 5. Input to program GETMODL, defining the model of LIN's et al.

a Input for the first run

```
MY + G(J) + G(K) + S(J,K) + E(L) ;
NE = < number of environments > ; NG = < number of genotypes > ;
S#LIST = (J,K,L) ;
E(NE) = 0 ; DO L = 1 TO NE - 1 ; E(NE) = E(NE) - E(L) ; END ;
G(NG) = 0 ; DO J = 1 TO NG - 1 ; G(NG) = G(NG) - G(J) ; END ;
DO J = 2 TO NG ; DO K = 1 TO J - 1 ;
S(J,K) = S(K,J) ; END ; END ;
DO J = 1 TO NG ; X = 0 ; DO K = 1 TO NG ;
X = X + S(J,K) ; END ; X = X - S(J,J) ;
S(J,J) = - 0.5 * X ; END ;
P#LIST = (MY) ;
DO J = 1 TO NG - 1 ; P#LIST = (P#LIST,G(J)) ; END ;
DO L = 1 TO NE - 1 ; P#LIST = (P#LIST,E(L)) ; END ;
DO J = 1 TO NG - 1 ; DO K = J + 1 TO NG ;
P#LIST = (P#LIST,S(J,K)) ; END ; END ;
P#SAVE = 1 ; NTOT = < number of data > ; %
```

b Input for the second run

```
MY + G(J) + G(K) + S(J,K) + E(L) + (BG(J) + BG(K) + BS(J,K)) * EDACH(L) ;
now are following the same statements as in the first run, excepted »P#SAVE = 1«, the con-
straints on BG and BS are defined in the same way as those for G and S, and the BG (*) and BS (
(*,*) will be entered into the CHAIN-variable P#LIST also like G (*) and BG (*,*);
moreover, we have:
EDACH(NE) = 0 ; DO L = 1 TO NE - 1 ;
EDACH(L) = #P#(NG+L) ; EDACH(NE) = EDACH(NE) - EDACH(L) ; END ; %
```

Statistical analysis of a generalized linear model can be performed by the following steps:
1) The restricted model: $E(\vec{y}) = T \cdot \vec{b}$, is passed to the program system, the corresponding design matrix T is generated, and $\hat{\vec{b}}$ is calculated.
2) The values of the coefficients of $\hat{\vec{b}}^r$ are passed to the FORMAC-system
3) The formal model: $E(\vec{y}) = T \cdot \vec{b} + \hat{U} \cdot \vec{a}$ is passed to the system, the corresponding design matrix (T : U) is generated, and the formal design parameters are calculated.
4) The hypothesis $\vec{a} = \vec{0}$ is tested.

The usage of our program system for this kind of analysis is demonstrated using a model which describes the genotype-environment-interactions in a diallel analysis, worked out by LIN et al. (1977). The random variable $y_i = y_{j,k,l}$ is the value of a metric trait, which was achieved in a cross of genotypes j and k in environment l. As in the model of MANDEL's, the genotype-environment interactions are assumed to be linear dependent on the genotypic effects:

$$y_i = y_{j,k,l} \sim N(\mu + g_j + g_k + s_{j,k} + \varepsilon_l + \left[\beta_{(g)j} + \beta_{(g)k} + \beta_{(s)j,k}\right] \cdot \varepsilon_l, \sigma^2)$$

The parameters g_* are the genotypic effects and the $s_{j,k}$ the interactions between them. $\beta_{(g)*}$ and $\beta_{(s)j,k}$ are the regression coefficients of the g_* and $s_{j,k}$, resp., on the environmental effects ε_l. The following constraints are imposed on the parameters:

$$\sum g = \sum \varepsilon = \sum \beta_{(g)} = 0, s_{j,k} = s_{k,j}, \beta_{(s)j,k} = \beta_{(s)k,j}, s_{j,j} = -\sum_k s_{j,k}, \beta_{(s)j,j} = -\sum_k \beta_{(s)j,k}.$$

To facilitate the treatment of a generalized linear model, we will suppose that the statement: »P#SAVE = 1;« occuring in the input to GETMODL, will cause the invoking program to store the values of the design parameters in the indexed symbolic variable #P# (cf. chapter 7). The input to GETMODL for an analysis of LIN's et al. model is given in table 5.

6.4 Survey on non-normal models

Design matrices do not only occur in connection with normal distributed homoskedastic errors. An important field of application is the analysis of quantal response data, where the parameters π_i of a binomial or λ_i of a Poisson distributed variable is obtained by: π_i or $\lambda_i = f(\vec{s}_{i,*}^{\,T} \cdot \vec{p})$, f being for instance a logit, loglog, probit or arc sine transformation. The statistical analysis of quantal response data is extensively described by LINDER and BERCHTOLD (1976).
NELDER (1968) suggests a model with two transformations, one to normalize the error and one to get the expectation from a linear model. So he gets: $y_i^{q1} \sim N((\vec{s}_{i,*}^{\,T} \cdot \vec{p})^{q2}, \sigma^2)$, where the parameters q_1 and q_2 have to be estimated, together with the design parameters, by an iterative algorithm.
A distribution free method for estimating design parameters which is based on rank procedures, has been published by JAECKEL (1972) and improved by MCKEAN and HETTMANSPERGER (1978),corresponding hypothesis tests are described by MCKEAN and HETTMANSPERGER (1976).

7. The Invoking Program LINMOD

The four formula manipulating subprocedures GETMODL, GENDES, GETHYP and GENHYP can be invoked by any PL/I- or FORTRAN-program which performs a numerical evaluation of a statistical model, in which a design matrix is involved. In order to facilitate the access to our program system, we have written a main program LINMOD, which carries out an analysis of the general linear model with fixed effects (see chapter 6.1) as well as for the generalized linear model defined in chapter 6.3, making use of the four formula manipulating subprocedures. LINMOD performs parameters estimation and calculates the sums of cross products of errors, as well as the sums of cross products due to a linear hypothesis. If the observation variable is univariate, an F-test is carried out, simultaneous confidence intervals are calculated for those parameter combinations by which the hypothesis was defined, as well as for additional contrasts given in @#LIST to subprogram GETHYP. Moreover, linear subhypotheses are tested as described in chapter 6.1. At two places, the program invokes optional user written subprograms to which information is passed concerning the matrices, vectors, and scalars evaluated so far. Using this interface ability, the user could, for instance, perform a test of hypothesis in a multivariate design, making use of the matrices of sums of cross products which are calculated by LINMOD and passed to the user program. When an analysis is acomplished, the assignments to the symbolic parameters are destroyed in the FORMAC-system and the input for the next run is immediately read in.
LINMOD can interprete four optional switch variables, which are passed by the input to GETMODL: 1) If S#COOK = 1, then LINMOD calculates, for each variate of Y, the studentized residuals t_i, to be used in a test of outliers, and the influence which every $y_{i,j}$ has on $\hat{\vec{p}}_{*,j}$ (see chapter 6.1). The calculation of these values in a linear design has been proposed by COOK (1977). 2) By default, parameter estimation is performed by calculating $\hat{P} = (S^TS)^{-1} S^T \cdot Y^T$. If ESTPAR = 2, then an orthonormalization procedure is used to minimize $(Y - SP)^T (Y - SP)$. 3) If P#SAVE = 1 then the values of $\hat{p}$ will be stored in the indexed symbolic variable #P# (cf chapter 6.3). 4) If GO#ON = 1, then the observation data of the current analysis will be reused in the next run, such that the same data set is analysed making use of two different models. This is necessary

Table 6. Flow of control in program LINMOD
Names of FORMAC-variables are written in italics

A: Invoke GETMODL. GETMODL reads information about the underlying model and passes it to the FORMAC-system, furthermore it passes the number of data, of design indices and of design parameters to the invoking program
B: If *GO# ON* − 1 then NREAD ← 1, otherwise NREAD ← 0
C: Allocate observation matrix and matrix of design parameters
D: Go to H
E. If NREAD = 0 then go to A
F: NREAD ← 2
G: Invoke GETMODL
H: Allocate design matrix
I: Invoke GENDES. GENDES builds up the design matrix using the information read in by GETMODL; if NREAD < 2, then the observation data and design indices are read in, otherwise they are supposed to be already stored.
J: Print out design and observation matrices
K: Perform parameter estimation, if *ESTPAR* = 2 by orthonormalisation, otherwise by inverting S^TS
L: If *P# SAVE* = 1, then assign values of estimates to the FORMAC-variable *# P#* (*)
M: If *S# COOK* = 1, then perform test for outliers
N: Invoke GETHYP. GETHYP reads information about the test to be performed and passes it to the FORMAC-system, furthermore it passes the number of lines of the hypothesis matrix, the number of contrasts to be constructed and the number of subhypotheses to be tested to the invoking program
O: If no hypothesis remains to be tested, then go to E
P: Allocate hypothesis matrix, result matrix, contrast matrix, matrix of results of subhypotheses
Q: Invoke GENHYP. GENHYP builds up the matrices which have been allocated in step P, using the information read in by GETHYP
R: Calculate and print sums of squares of hypothesis
S: If analysis is multivariate, then go to N
T: Calculate F-value and Scheffé-intervals, test subhypotheses
U: Go to N

The program stops when a predefined symbol is read in.

in the case of the generalized linear model (chapter 6.3), and, moreover, in the case of the general linear model when $(S^TS)^{-1}$ cannot be obtained and SP_{hyp} is to be calculated by explicitely formulating the restricted model as it is imposed by H_o (see chapter 6.1). The flow of control in program LINMOD is given in table 6.

References

ANDREWS, D. F. (1971): Significance tests based on residuals. Biometrika *58*, 139–148

ANSCOMBE, F. J. and J. W. TUKEY (1963): The examination and analysis of residuals. Technometrics *5*, 141–160

BAHR, K. A. (1977): FORMAC73 USER's Manual. GMD Darmstadt

BEHNKEN, W. D. and N. R. DRAPER (1972): Residuals and their variance patterns. Technometrics *14*, 101–111

BERCHTOLD, W. (1977): Lineares Modell, Schätzbarkeit und Computer. EDV in Medizin und Biologie *8*, 129–134

BOWDEN, D. C. (1970): Simultaneous confidence bands for linear regression models. JASA *65*, 413–421

BOWMAN, K. O. (1973): Power of the kurtosis statistics, b_2, in tests of departures from normality. Biometrika *60*, 623–628

BRYCE, G. R. and CARTER, W. (1974) MAD – The analysis of variance in unbalanced designs – a software package. In: COMPSTAT 1974, Proceedings in computational statistics, Physika Verlag, Wien 1974, 447–456

BULMER, M. G. (1958): Confidence intervals for distance in the analysis of variance. Biometrika *45*, 360–369

COOK, R. D. (1977) Detection of influential observation in linear regression. Technometrics *19*, 15–18

CORBEIL, R. R. and R. S. SEARLE (1976a): Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics *18*, 31–38

–,– (1976b): A comparison of variance component estimates. Biometrics *32*, 779–791

CUNNINGHAM, E. P. and C. R. HENDERSON (1968): An iterative procedure for estimating fixed effects and variance components in mixed model situations. Biometrics *24*, 13-25

D'AGOSTINO, R. and E. S. PEARSON (1973): Tests for departure from normality. Empirical results for the distribution of b_2 and $\sqrt{b_1}$. Biometrika *60*, 613–622

DEMPFLE, L., G. HEIL and K. RUTZMOSER (1977): Relative Effizienz verschiedener Methoden zur Schätzung von Varianzkomponenten. EDV in Medizin und Biologie *8*, 52–56

ELLENBERG, J. H. (1973): The joint distribution of the standardized least squares residuals from a general linear regression. JASA *68*, 941–943

FISHER, R. A. (1918): The correlations between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinburgh (Part II) *52*, 399–433

FORKMANN, G. and W. SEYFFERT (1977): Simulation of quantitative characters by genes with biochemically definable action. VI. Modifications of a simple model. Genetics *85*, 557–572

GEBHARDT, F. (1966): Verteilung und Signifikanzschranken des 3. und 4. Stichprobenmoments bei normalverteilten Variablen. Biom. Z. *8*, 219–241

GEIDEL, H. and M. PRECHT (1978): Zur numerischen Genauigkeit von Regressionsprogrammen. EDV in Medizin und Biologie *9*, 77–84

HARDWICK, R. C. and J. T. WOOD (1972): Regression methods for studying genotype-environment interactions. Heredity *28*, 209–222

HARTLEY, H. O. and J. N. K. RAO (1967): Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika *54*, 93–108

HAYMAN, B. I. (1954): The analysis of variance of diallel tables. Biometrics *10*, 235–244

HEMMERLE, W. J. and H. O. HARTLEY (1973): Computing maximum likelihood estimates for the mixed A. O. V. model using the W transform. Technometrics *15*, 819-831

HENDERSON, C. R. (1953): Estimation and covariance components. Biometrics *9*, 226–252

IBM-Corp (1967): PL/I-FORMAC Interpreter. New York: Hawthorne

JAECKEL, L. A. (1972): Estimating regression coefficients by minimizing the dispersion of the residuals. Ann. Math. Statist. *43*, 1449–1458

JENSEN, D. R. and M. Q. JONES (1969): Simultaneous confidence intervals for variances. JASA *64*, 324–332

KLEIN, H. D. and H. D. QUEDNAU (1976): Die biometrische Bewertung des Univalenten-Verhaltens in konjugationsgestörten Mutanten von Pisum sativum. TAG *48*, 227–235

KRES, H. (1975): Statistische Tafeln zur multivariaten Analyse. Springer-Verlag Berlin, Heidelberg, New York

LIN, C. S., M. R. BINNS and B. K. THOMPSON (1977): The use of regressing methods to study genotype-environment interactions. Heredity *38*, 309–319

LINDER, A. and W. BERCHTOLD (1976): Statistische Auswertung von Prozentzahlen. UTB, Band 522 (Birkhäuser, Basel)

LUND, R. E. (1975): Tables for an approximate test for outliers in linear models, Technometrics *17*, 473–477

MANDEL, J. (1961): Non-additivity in two-way analysis of variance. JASA *56*, 878–888

MCKEAN, J. W. and T. P. HETTMANSPERGER (1976): Tests of hypothesis based on ranks in the general linear model. Comm. Statist. A *5*, 693–709

–,– (1978): A robust analysis of the general linear model based on one step R-estimates. Biometrika *65*, 571–579

MILLIKEN, G. A. and F. A. GRAYBILL (1970): Extensions of the general linear hypothesis model. JASA *65*, 797–807

NELDER, J. A. (1968): Weighted regression, quantal response data, and inverse polynomials. Biometrics *24*, 979–985

NOLLAU, W. (1975): Ein Verfahren zur Kovarianzanalyse bei ungleicher Zellenbesetzung, Teil 1: Das Modell mit fixen Effekten. EDV in Medizin und Biologie *6*, 104–110

–,– (1976): Ein Verfahren zur Kovarianzanalyse bei ungleicher Zellenbesetzung, Teil 2: Das gemischte Modell. EDV in Medizin und Biologie *7*, 16–21

NORDSTRÖM, M., E. SANDEWALL and D. BRESLAW (1973): LISPF1: FORTRAN implementation of LISP 1.5. University of Uppsala, Department of Computer Sciences

NOWAK, H. (1975): Universalprogramme zur Auswertung linearer Modelle. Statistical Software Newsletter *3*, 42–47

PATNAIK, P. B. (1949): The non-central χ^2- and F-distributions and their applications. Biometrika *36*, 202–232

PATTERSON, H. D. and R. THOMPSON (1971): Recovery of interblock information when block sizes are unequal. Biometrika *58*, 545-554

PRECHT, M. (1978): Zur numerischen Lösung und zur Kondition von Regressionsproblemen. Statistical Software Newsletter *4*, 36–41

PRESCOTT, P. (1975): An approximate test for outliers in linear models. Technometrics *17*, 129–132

PUTTER, J. (1967): Orthonormal bases of error spaces and their use for investigating the normality and variance of residuals. JASA *62*, 1022–1036

QUEDNAU, H. D. (1973): MAXLIKE – Ein Programmsystem zur Parameterschätzung beliebiger Verteilungen nach der Maximum Likelihood-Methode. EDV in Medizin und Biologie *4*, 37-41

–,– (1975): Die Anwendung der Maximum Likelihood-Methode auf die eigentlich nichtlineare Regressionsanalyse. Biom. Z. *17*, 225–231

–,– (1976a): LIBAFORM – Eine Computersprache zur symbolischen Verarbeitung mathematischer Formeln auf LISP-Basis. Applied Informatics *18*, 168–174

–,– H. D. (1976b): The comparison of parameters estimated from several different samples by maximum likelihood. Biometrics *32*, 683–688

–,– (1976c): Maximum Likelihood – Schätzungen und Likelihood Quotiententests bei nicht normalverteiltem, zweiwegkreuzklassifiziertem Datenmaterial mit festen Effekten. EDV in Medizin und Biologie *7*, 87–91

–,– (1979): Generating Programs in high level programming languages by LIBAFORM, in preparation

RAO, C. R. (1962): A note on generalized inverse of a matrix with applications to problems in mathematical statistics. J. Roy. Stat. Soc., Series B, *24*, 152–158

–,– (1972): Estimation of variance and covariance components in linear models. JASA *67*, 112–115

RUST, B., W. R. BURRUS and C. SCHNEEBERGER (1966): A simple algorithm for computing the generalized inverse of a matrix. Comm. ACM *9*, 381–385

SCHEFFE, H. (1953): A method for judging all contrasts in the analysis of variance. Biometrika *40*, 97–104

–,– (1959): The Analysis of Variance. John Wiley & Sons, Inc., New York

SEARLE, S. R. (1968): Another look at Henderson's method of estimating variance components. Biometrics *24*, 749–787

SPRENT, P. (1969): Models in regression and related topics. Methuen, London

THOMPSON, R. (1975): A note on the W transform. Technometrics *17*, 511–512

TOPHAM, P. B. (1966) Diallel analysis involving maternal and maternal interaction effects. Heredity *21*, 665–674

TUKEY, J. W. (1949): One degree of freedom for non-additivity. Biometrics *5*, 232-242

URFER, W. and H. THÖNI (1976): Zur Schätzung von Wachstumskurven aufgrund wiederholter Messungen am gleichen Individuum. EDV in Medizin und Biologie *7*, 92–95

WALTERS, D. E. and J. S. GALE (1977): A note on the Hayman analysis of variance for a full diallel table. Heredity *38*, 401–407

WATKINS, R. and L. P. S. SPANGELO (1968): Components of genetic variance in the cultivated strawberry. Genetics *59*, 93–103

WEARDEN, S. (1964): Alternative analyses of the diallel cross. Heredity *19*, 669–680

WEILING, F. and C. UNGER (1977): Über einige praktische Erfahrungen zur Leistungsfähigkeit der verschiedenen, in der multivariaten Varianzanalyse (MANOVA) verwendeten Testverfahren. Biom. J. *19*, 549–559

WILKINSON, G. N. and C. E. ROGERS (1973): Symbolic description of factorial models for analysis of variance. J. Roy. Stat. Soc. C., Applied Statistics *22*, 392–399

WOTTAWA, H. (1974): Das ›allgemeine lineare Modell‹ – Ein universelles Auswertungsverfahren. EDV in Medizin und Biologie *5*, 65–73

YATES, F. and W. G. COCHRAN (1938): The analysis of groups of experiments. J. agric. Sci. Camb. *28*, 556–580

Anschrift des Verfassers: Dr. H. D. Quednau, Institut für Genetik der Universität Bonn, Kirschalle 1, D-5300 Bonn

EDV in Medizin und Biologie **10** (3), S. 80–82, ISSN 0300-8282

Ein Microcomputersystem zur direkten (morphometrischen) Bestimmung von Bakterienbiomasse in der Limnologie

H.-J. Krambeck, Christiane Krambeck und J. Overbeck

Zusammenfassung

Die exakte Bestimmung der Bakterienbiomasse ist eine der mühsamsten Aufgaben in der Limnologie, weil es bisher keine automatischen Bildanalysegeräte gibt, die zwischen den Bakterien und anderen im Wasser vorhandenen Partikeln (Detritus) hinreichend gut unterscheiden können.

Die hier vorgelegte Methode verbindet daher die Überlegenheit des Menschen, komplexe Formen zu erkennen mit der Überlegenheit des Computers, alle notwendigen Berechnungen schnell und fehlerlos durchzuführen. Das von den Autoren entwickelte System besteht aus einem M 6502-Mikrocomputer mit 16 K Bytes-Speicher, einem Summagraphics Digitizer und einem Decwriter als Dialogterminal. Das in Focal geschriebene Programm erlaubt den Dialog zwischen Benutzer und Rechner und reduziert die Arbeit der exakten (morphometrischen) Bestimmung von Bakterienbiomasse in etwa auf die des reinen Zählens der Zellen; zudem testet es, ob ein Bakterium schon gemessen wurde und führt auch die notwendigen statistischen Berechnungen (T-Test, Vertrauensbereiche) durch.

Der modulare Aufbau des in der Sprache Focal geschriebenen Programms erleichtert die Anpassung an andere Fragestellungen, etwa die Bestimmung des Volumens von fädigen Algen.

Summary

The exact estimation of bacterial biomass is one of the most striving and errorprone tasks in limnology, yet all available automatic picture analyzers failed up to now in this field. The presented method therefore combines the superiority of men to process complex patterns with the superiority of computers to do all necessary calculations. The computersystem developed by the authors comprises a M 6502 microcomputer with 16 K bytes memory, a summagraphices digitizer and a decwriter. The program written in Focal allows for a dialogue between user and computer; it reduces the work of the exact (morphometric) estimation of the bacterial biomass to about the work of just counting the number of cells, also all checking and the necessary statistics are included. Furthermore the modular structure of the high-level language program alleviates the adaption to different problems, e.g. the estimation of the biomass of algae clones.

1. Einleitung und Problematik

Die Kenntnisse der Verteilung der bakteriellen Biomasse in der Wassersäule von Binnengewässern ist ein zentrales Anliegen des Limnologen (OVERBECK 1972).

Die anfallenden Planktonproben enthalten bei geringen Zelldichten ($\approx 10^5$ Bakterien/ml) zahlreiche andere Organismen (vor allem Algen) und amorphe Partikel (Detritus), die sich nicht quantitativ von den Bakterien trennen lassen. Daher sind chemisch-physikalische Methoden wie Trockengewichts-, Eiweiß- oder Trübungsbestimmungen unbrauchbar. Die Bakterienbiomasse muß direkt mikroskopisch analysiert werden. Wegen der schwankenden Größenverteilung der Bakterienpopulation sind auch die üblichen Zellzählungen unbefriedigend und nur Volumenbestimmungen liefern Werte, die der Biomasse der Bakterienpopulation im Gewässer proportional sind. Die Menge der dabei anfallenden Meßdaten (400–1000/Probe) hat die Methode jedoch bisher auf einige exemplarische Untersuchungen beschränkt (KRAMBECK 1978).

Daher wurde das hier vorgestellte System entwickelt, um diese in der Limnologie so wichtige Methode soweit zu verbessern, daß der Arbeitsaufwand in etwa auf den des reinen Zählens der Zellen reduziert war.

Die in Abb. 1 dargestellte Rasterelektronenmikroskop-Aufnahme einer Probe von Gewässerbakterien macht zwei Eigentümlichkeiten der Bearbeitung dieser Art von Proben deutlich: Es erfordert einerseits Erfahrung, die eigentlichen Bakterien auf dem Foto zu erkennen und die in unserem Institut bisher vorgestellten automatischen Bildanalysegeräte haben de facto vollkommen versagt; auf der anderen Seite ist die Volumenberechnung des erkannten Bakteriums ziemlich einfach. Trotzdem war bei allen uns bekannten Bildanalyse-Systemen (auch den nicht vollautomatischen) keine passende Software zu finden.

Die optimale Lösung des Problems lag für uns darin, ein halbautomatisches System zu benutzen, also dem menschlichen Benutzer den Teil zu überlassen, in dem er dem Computer weit überlegen ist, nämlich das Erkennen von komplexen Formen, d. h. hier das Identifizieren der Bakterien, und an einem nachgeschalteten, frei programmierbaren Rechner ein maßgeschneidertes Programm zu entwickeln, das alle (!) übrigen Arbeiten erledigen sollte.

2. Methodik und Aufgaben für das System

Gewässerbakterien haben in der Regel eine langgestreckte, rotationssymmetrische Form (Abb. 1), deren Volumen sich sehr gut durch einen Zylinder mit zwei aufgesetzten Halbkugeln approximieren läßt (Abb. 2). Nach der Präparation liegen sie auf ebenen Filteroberflächen und werden genau von oben fotografiert, so daß ihre Länge und Breite direkt gemessen werden können.

Dabei wird überprüft, ob das Scalarprodukt der Vektoren 1 und b sehr klein, φ also etwa 90° ist. Außerdem wird getestet,

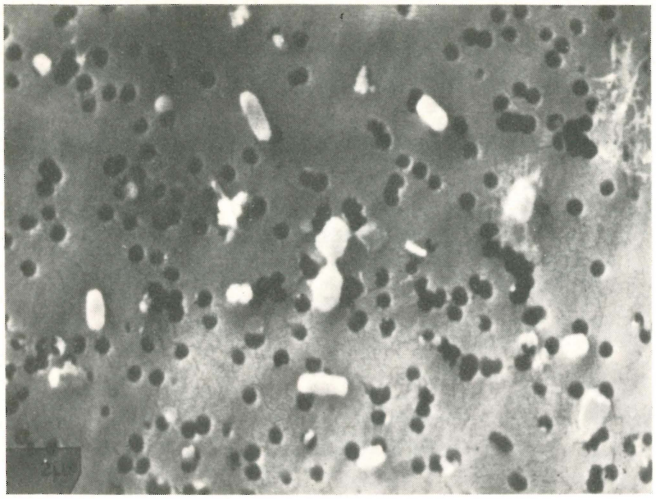


Abb. 1. Eine rasterelektronenmikroskopische Aufnahme der im Text bearbeiteten Bakterienprobe.

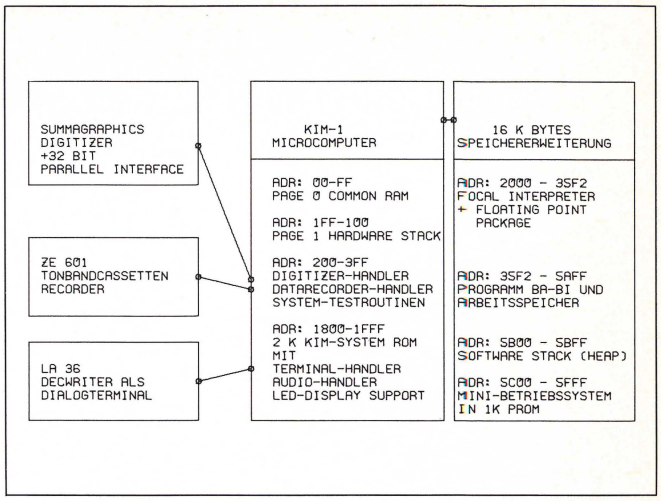


Abb. 3. Die Hardware- und Software-Komponenten sowie die Speicheraufteilung des Systems.

ob sich die Mittelpunktskoordinaten der Strecken l und b um nicht mehr als die Strecke d unterscheiden. Wenn nicht beide Kriterien erfüllt sind, verlangt der Rechner die nochmalige Vermessung des fraglichen Bakteriums, siehe auch den Dialog in Abb. 5.

Durch das Bestimmen der Punkte (x_i), y_i, $i = 1, 2, 3, 4$ mit einem Digitizer lassen sich l und b und daraus das Bakterienvolumen leicht bestimmen:

$$l = \sqrt{(x_3-x_1)^2 + (y_3-y_1)^2},$$
$$b = \sqrt{(x_4-x_2)^2 + (y_4-y_2)^2}$$

Das Gesamtvolumen setzt sich aus dem zylindrischen Teil

$$V_{zyl} = \pi \cdot (\frac{b}{2})^2 \cdot (l-b)$$

und dem kugelförmigen Teil

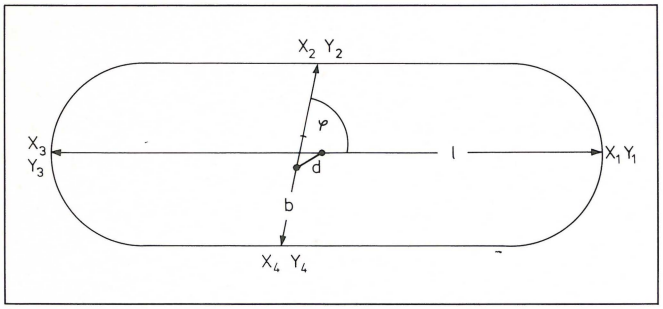
$$V_{Kugel} = \frac{4}{3} \cdot \pi \cdot (\frac{b}{2})^3$$

zusammen.

Rein kugelförmige Zellen sind in
$V_{ges} = V_{zyl} + V_{Kugel}$
mit $l = b$ und damit als Spezialfall
$V_{zyl} = 0$ enthalten.

Für den Bearbeiter der Bakterienprobe besteht also die Aufgabe lediglich darin, die Bakterien zu identifizieren und die vier Koordinatenpaare x_i, y_i $(i = 1, 2, 3, 4)$ jedes Bakteriums an den Rechner weiterzugeben.

Abb. 2. Die morphometrischen Daten des Bakteriums (x_i, y_i) $i = 1, 2, 3, 4$ sowie die Prüfgrößen φ und d.



Der Rechner hat die Aufgabe, im Dialog mit dem Bearbeiter für die gesamte Datenorganisation und Kontrolle zu sorgen. Dabei soll sich die Software »freundlich« benehmen, d. h. auf alle möglichen Benutzer-Fehler und -Wünsche eingehen und mit wachsender Vertrautheit des Benutzers kanpper im Dialogtext werden.

3. Hardware- und Software-Komponenten

Die Hardware des Systems besteht aus einem M 6502-Microcomputer in Form der KIM-1 Platine (Fa. Commodore, MOS Technology), einer 16 K Byte-Speichererweiterung (Fa. Astronik), einem ZE 601-Tonbandkasettenrecorder mit FSK-Modem zum Laden und Speichern der Software (Fa. Neumüller), einem LA 36 Decwriter als Dialogterminal (Fa. Digital Equipment) sowie einem Digitizer zur Bereitstellung der Bakterienkoordinaten (Fa. Summagraphics).

Die Software besteht aus den in Assembler geschriebenen Handlern für den Digitizer und den Datarecorder sowie einem kleinen Betriebssystem, das das Retten und Laden des Speicherinhalts gestattet. Diese Programme wurden in einem in Algol 60 geschriebenen und auf einer DEC PDP 8 inplementierten Crossassembler (KRAMBECK 1978) entwickelt; das Betriebssystem wurde dann in einen Prom übertragen, um die Bedienung des Systems möglichst zu vereinfachen.

Die weiteren Software-Komponenten sind: Der Focal-Interpreter (von Digital Equipment für die PDP 8 entwickelt, von »the 6502 Program Exchange« auf den M 6502 umgeschrieben) sowie das in Focal geschriebene Dialogprogramm BA-BI, das das in Abb. 5 wiedergegebene Flußdiagramm ausführt. Abb. 3 faßt die Hard- und Software-Komponenten sowie die Speicheraufteilung (hexadezimal) zusammen.

Die Sprache Focal wurde der in Microcomputern weit verbreiteten Sprache Basic vorgezogen, weil sie bedeutend höheren Programmierkomfort bietet, voll rekursiv ist und wesentlich geringeren Platzbedarf als ein gleichlanges Basicprogram hat; das liegt daran, daß Focal für einen Minicomputer (die PDP 8) konzipiert wurde, während Basic zuerst für einen Mainframe geschrieben wurde.

Ein kleines Beispiel für Focal enthält die Abb. 4.

```
*W

 C 6502-FOCAL HJK 21-MAY-79

 1.10 C RECURSIVE CONVERSION DECIMAL TO OCTAL
 1.20 A !!" DECIMAL: ",D; T " OCTAL  = "; D 1.3; G 1.2
 1.30 S Z(N=N+1)=D-8*D=FINT(D/8); ON (-D) 1.3; T Z(N); S N=N-1

*G

 DECIMAL: 4097
 OCTAL  = 10001

 DECIMAL: 511
 OCTAL  = 777

 DECIMAL:
KIM
2000 78
```

Abb. 4. Ein kleines Beispiel für die verwendete Programmsprache Focal.

4. Der Dialog zwischen Benutzer und Rechner

Das Dialogprogramm ist so konzipiert, daß die Benutzung der Anlage ohne spezielle Schulung möglich ist; die vom Programm gestellten Fragen sind für die Bediener der Anlage (zumeist biol.-technische Assistenten) einleuchtend und werden daher problemlos richtig beantwortet.

Abb. 5 zeigt das Flußdiagramm des Dialog-Programms BA-BI; Abb. 6 einen Originaldialog mit dem Benutzer des Systems.

5. Diskussion

Das beschriebene Microcomputersystem erlaubt die direkte morphometrische Volumenbestimmung von Bakterienzellen. Das Endergebnis (Bakterienbiomasse/Volumen Seewasser) wird jeweils mit Vertrauensbereichen (95%) und Probenbeschreibung ausgegeben. Darüber hinaus übernimmt der Rechner im Dialog mit dem Benutzer die gesamte Datenorganisation und Kontrolle, so daß der Zeitaufwand kaum größer ist als beim reinen Zählen der Zellen, während die Genauigkeit um etwa eine Größenordnung besser wird.

Da die Software in der Programmiersprache Focal geschrieben ist, läßt sich das Programm den wechselnden Fragestellungen der Planktonanalyse sehr einfach anpassen. Um z. B. zusätzlich Größenverteilungen anzugeben oder auch um Algenpopulationen zu untersuchen, brauchen nur einzelne Unterprogramme zugefügt oder geändert zu werden.

Literatur

KRAMBECK, H.-J. (1978): Ein PDP 8 (PDP 11) Programmentwicklungssystem für den MOS Technology (KIM 1) Microcomputer und seine Anwendung in einer limnologischen Meßanlage. – Zusammenfassung der Referate des ersten deutschsprachigen DECUS-Symposiums 3.–4. April 1978, S. 27–28.

KRAMBECK, Christiane (1978): Changes in planktonic microbial populations - an analysis by scanning electron microscopy. - Verh. Internat. Verein. Limnol. *20*, 2255–2259.

OVERBECK, J. (1972): Experimentelle Untersuchungen zur Bestimmung der bakteriellen Produktion im See. - Verh. Internat. Verein. Limnol. *18*, 176–187.

Eingegangen: 15. 6. 1979

Anschrift der Verfasser: Dr. Hans-Jürgen Krambeck, Dr. Christiane Krambeck, Prof. Dr. Jürgen Overbeck, Max-Planck-Institut für Limnologie, Abteilung Allgemeine Limnologie, August-Thienemann-Str. 2, D-2320 Plön.

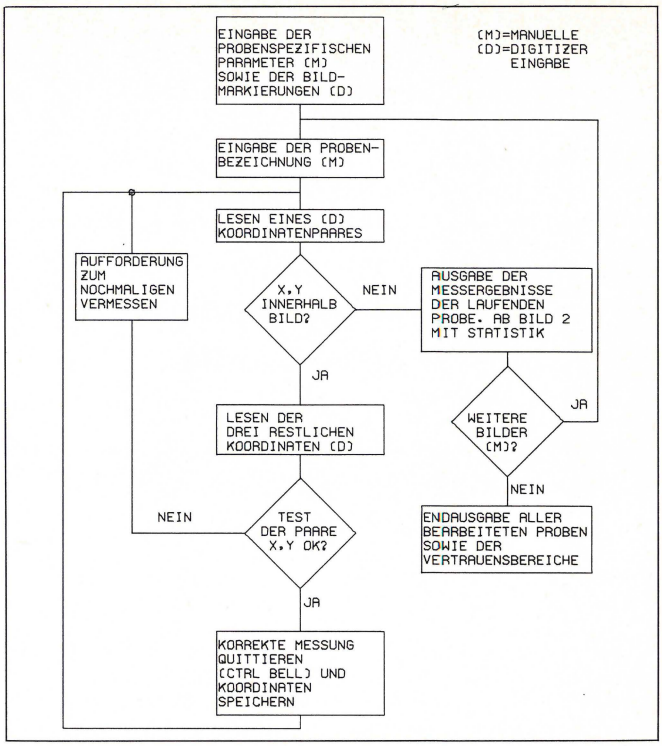


Abb. 5. Das (vereinfachte) Flußdiagramm des Dialogprogramms BA-BI.

Abb. 6. Der Auswertungsgang als Dialog zwischen Anwender und Microcomputer; die Angaben des Anwenders sind unterstrichen.

```
*G

BAKTERIEN-BIOMASSE UND -ZELLZAHL. VERSION FEBR-1979.

FILTERDURCHMESSER [MM] (Z.B.: 22.1 UND 'RETURN') =26.5
LAENGE DER MICRON-MARKE [MICRON] (Z.B.: 2 UND 'RETURN') =2
ENDPUNKTE DER MICRON-MARKE AUF PROJIZIERTEM BILD MESSEN
(2 MESSUNGEN):  LAENGE= 19[MM]
DIGONALE ECKEN DES PROJIZIERTEN BILDES MESSEN:  FORMAT=250X269[MM^2]
FORTSETZUNG VON FRUEHER BEARBEITETER PROBE? JA ODER NEIN:NEIN.


TEXT ZU  1. PROBE
(MIT 'RETURN' BEENDEN.)

TEST-MESSUNGEN ZU KRAMBECK ET AL. 1979


PROBENVOLUMEN[ML] (+'RETURN') =50

DOPPELTEST UND BILDAUSDRUCK?JA ODER NEIN:NEIN.


  1. BILD. NEGATIV NR. (+'RETURN'): FILM 7E1

BAKTERIENMESSUNGEN:
BAKTERIUM NOCHMAL MESSEN:
BAKTERIUM NOCHMAL MESSEN:
BILD LOESCHEN?JA ODER NEIN:NEIN.

ANZAHL AUF BILD  1=  8
MITTLERES BAKTERIENVOLUMEN [MICRON^3] =      5.315


TEST-MESSUNGEN ZU KRAMBECK ET AL. 1979
      8 BAKTERIEN AUF  1 BILD(ERN) GEMESSEN.
BIOMASSE [10^6 MICRON^3 / ML] =    0.612
ZELLZAHL [10^6/ML]            =    0.115

WEITERE BILDER ZU LAUFENDER PROBE?JA ODER NEIN:JA.


  2. BILD. NEGATIV NR. (+'RETURN'): FILM 7E2

BAKTERIENMESSUNGEN:
BILD LOESCHEN?JA ODER NEIN:NEIN.

ANZAHL AUF BILD  2= 10
MITTLERES BAKTERIENVOLUMEN [MICRON^3] =      4.422


TEST-MESSUNGEN ZU KRAMBECK ET AL. 1979
     18 BAKTERIEN AUF  2 BILD(ERN) GEMESSEN.
BIOMASSE [10^6 MICRON^3 / ML] =    0.624   +-   0.156
ZELLZAHL [10^6/ML]            =    0.129   +-   0.183

G=2 A4=18 A5=100000 B4=45852 B5=2038
WEITERE BILDER ZU LAUFENDER PROBE?JA ODER NEIN:NEIN.

JA ODER!
```

EDV in Medizin und Biologie **10** (3) S. 83–85, ISSN 0300-8282

The use of the principal component regression analysis in structure-activity studies*

P. P. Mager**

Summary

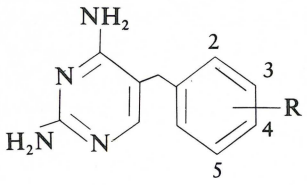
The principal component regression analysis is used when the predictor variables correlate significantly with the regressor variables but when a collinearity and multicollinearity exists among the regressors.

Zusammenfassung

Die Hauptkomponenten-Regressionsanalyse kann dann bei Struktur-Wirkungs-Studien benutzt werden, wenn die vorherzusagenden Variablen mit den Regressorvariablen signifikant korrelieren und wenn letztere kollinear und multikollinear sind. Im Gegensatz zum Reduktionsmodell der Regressionsanalyse tritt in diesem Fall kein Informationsverlust auf.

Introduction

The lack of selectivity in the action of many metabolic inhibitors is inherent in their mechanism of action due to the universality of biochemical mechanisms. Hitchings, using dihydrofolic acid reductase of bacterial and mammalian origin, has developed inhibitors on the basis of species differences (Hansch et al. [1977]). Considering the molar concentration for 50 % inhibition of dihydrofolic acid reductase of *E. coli*, expressed in terms of $Y_1 = -ln\ I_{50}$, and the relative potency in comparison to the 50 % inhibition of dihydrofolic acid reductase of rat liver, expressed in terms of $Y_2 = ln\ (I_{50}$ of rat liver$/I_{50}$ of *E. coli*), it can be visually assumed (Table 2) that the biological response of the substituents R of the basic molecule



depends on the following chemical constants: the dummy variable $X_1 = X_D$ which accounts for the number of hydrogen acceptor substituents (such as OMe) where the nonhydrogen bonders (Me, Cl) and the amphiprotic substituents (OH) are parameterized by zero, $X_2 = \sigma_R$ = sum of the resonance constants, and $X_3 = E_s$ = sum of Taft's steric constants, see literature (Mager [1976]). On the other hand, it can be also seen (Table 1) that there is a remarkable collinearity among the regressor variables and, perhaps, a considerable multicollinearity. In such cases, the regressors involved in multicollinearities tend to have small test statistics, in general (Kendall [1976]). In order to avoid this problem in multiple structure-activity regression analysis (Hansch [1971]) and multivariate structure-activity relationships (Mager [1979]), the principal component regression analysis was proposed (Mager [1979, 1980]).

Statistical analysis

First, we compute the correlation matrix **R** written in the partioned form. It can be seen that all correlation coefficients are significantly different from zero at the 1 % level or less (critical values of the population: 0.898 at the 0.1 % level, 0.798 at the 1 % level).

Table 1. Substituents and chemical constants of benzylpyrimidines, and first principal component function scores used in the Masca model (multivariate structure-activity analysis in combination with the multivariate bioassay).

Compd	R	X_1	X_2	X_3	Z_1
1	4-Me	0	−0.14	0.00	0.080388
2	4-Cl	0	−0.24	0.18	0.243516
3	4-OH	0	−0.66	0.69	0.784186
4	4-OMe	1	−0.55	0.99	1.467654
5	3-OMe	1	−0.55	0.99	1.467654
6	3.4-diOMe	2	−1.10	1.98	2.935308
7	3,4-diOMe,5-Cl	2	−1.34	2.16	3.178824
8	3,5-diOMe,4-OH	2	−1.76	2.67	3.719493
9*	3,4,5-triOMe	3	−1.65	2.97	4.402962

* Trimethoprim, the best compound and clinically used.

*) Lecture hold at the 5th Colloquy of the G.D.R. region of the International Biometric Society and of the Society of Physical and Mathematical Biology, March 20–23, 1979, Eisenach/Thuringia(G.D.R.).

**) Section of Pharmacy, Martin Luther University Halle-Wittenberg, G.D.R. – 402 Halle, Weinbergweg

We obtained that $\mathbf{R}$ =

$$= \left[\begin{array}{c|c} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \hline \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{array}\right]$$

$$= \left[\begin{array}{cc|ccc} 1 & 0.979793 & 0.950542 & -0.965925 & 0.978902 \\ & 1 & 0.931127 & -0.958506 & 0.966386 \\ \hline & & 1 & -0.897627 & 0.963508 \\ & & & 1 & -0.981344 \\ & & & & 1 \end{array}\right]$$

Experimentalists not familiar with such problems would assume that a combination of all chemical constants improves the correlation because each regressor variable is highly correlated with the both predicting variables. This total regression model is going out from $\mathbf{Y} = \mathbf{BX}'$ where $\mathbf{Y}' = (Y_1\ Y_2)$, $\mathbf{X}' = (X_1\ X_2\ X_3)$, and $\mathbf{B}$ is the regression matrix with the vector $\mathbf{b}_0$ of the intercepts. The analysis yields that $\mathbf{B}$ =

$$\begin{bmatrix} 11.7776 & 3.22798 & -8.22895 & -5.47391 \\ 2.3866 & 3.20402 & -8.77383 & -5.62124 \end{bmatrix}$$

with the multiple correlation coefficients $R_1 = 0.99001$ and $R_2 = 0.97713$. Although the both multiple correlation coefficients are significant at the 1% level or less, it is remarkable that only $b_{12} = -8.22895$ is significant at the 5% level, the other partial regression coefficients are nonsignificant even at the 5% level. The theoretically calculated values are listed in Table 2 (run 1).

The reason of this discrepancy between the overall-correlation significance and the significance of regression coefficients is the high collinearity (see $\mathbf{R}_{xx}$) and multicollinearity. The multicollinearity can be mathematically expressed by the internal coefficient of determination,

$D_i = 1 - 1/r^{ii}$ $\quad (i = 1, 2, \ldots$ regressors)

where r^{ii} is the main diagonal element of $\mathbf{R}_{xx}^{-1}$ =

$$\begin{bmatrix} 104.539503 & -135.484641 & -233.681687 \\ & 202.643350 & 329.403371 \\ & & 549.412196 \end{bmatrix}$$

We obtained that $D_1 = 0.99044$, $D_2 = 0.99587$, $D_3 = 0.99819$ (significant at the 1% level or less). In such cases, the simultaneous test statistics on regression coefficients (multiple regression implies that another test statistic must be used, the t test) tend to have small test statistics of the variables involved predominantly in multicollinearities, and these regressors are eliminated when the backward elimination is applied. Note that this technique allows that only significant regressors (at the 5% level or less) are included in the final regression matrix (Mager [1979]). The principal component analysis (Princo) is going out from

$det(\mathbf{R}_{xx} - \lambda_i \mathbf{E}) = 0$

where $\mathbf{E}$ is the unit matrix. The vector of the eigenvalues is

λ = (2.895436 0.103383 0.001191)'

(cum percent trace: 96.515%, 99.956%, 100%). The smallest eigenvalue tends to zero (another indication of multicollinearity), and the largest elements of the corresponding eigenvector $\mathbf{v}_3$ obtained from $(\mathbf{R}_{xx} - \lambda_i \mathbf{E})\mathbf{v}_i - 0$ show which variables are predominantly involved in multicollinearities. We obtained that

$\mathbf{V}$ =

$$\begin{bmatrix} 0.570451 & 0.746729 & -0.342026 \\ -0.574201 & 0.660336 & 0.483994 \\ 0.587265 & -0.679703 & 0.805461 \end{bmatrix}$$

Therefore, the variable X_3 will be predominantly eliminated in the reduced regression model (the large value of 0.805461 indicates this fact). The reduced regression model leads to $\mathbf{B}$ =

$$\begin{bmatrix} 12.15770 & 0.93352 & -2.30686 & 0 \\ 2.57608 & 0 & -4.08577 & 0 \end{bmatrix}$$

where the multiple correlation coefficients $R_1 = 0.98433$, $R_2 = 0.95850$ are significant at the 1% level or less, and all non-zero regression coefficients are significant at the 5% level or less ($b_{11} = 0.93352$ at the 5% level, $b_{12} = -2.30686$ at the 2% level, b_{22} at the 0.01% level). The internal determination coefficient is here $D = r^2(X_1, X_2) = 0.80573$ (significant at the 5% level). The theoretically calculated values are listed in Table 2 (run 2).

So far as one concerns the goodness of fit, as measured by the multiple correlations, the total and reduced model are here about as good as one another. Clearly, no precise meaning can be assigned to the individual regression coefficients because the correlation coefficients of $\mathbf{R}_{yx}$ are highly significant. This implies that the recognition (comparison between the experimentally observed and theoretically calculated data basing on the training set) must be carefully distinguished from the pre-

Table 2. Biological response variables obtained experimentally and calculated theoretically. Run 1 is based on the total and run 2 is based on the reduced model of multivariate regression, run 3 is based on the principal component regression analysis.

Compd	Obtd	Calcd Run 1	Calcd Run 2	Calcd Rund 3	Obtd	Calcd Run 1	Calcd Run 2	Calcd Run 3
	Y_1	Y_{1t}	Y_{1t}	Y_{1t}	Y_2	Y_{2t}	Y_{2t}	Y_{2t}
1	12.98	19.23	12.48	12.37	3.77	3.61	3.15	3.15
2	13.07	12.77	12.71	12.62	3.75	3.48	3.65	3.41
3	13.23	13.43	13.68	13.43	4.49	4.30	5.27	4.25
4	13.72	14.11	14.36	14.45	5.04	4.85	4.82	5.32
5	14.53	14.11	14.36	14.45	4.47	4.85	4.82	5.32
6	16.12	16.45	16.56	16.64	6.55	7.32	7.07	7.61
7	17.03	17.44	17.11	17.00	7.82	8.41	8.05	8.00
8	18.33	18.10	18.08	17.81	9.12	9.23	9.77	8.84
9	19.11	18.78	18.76	19.83	10.82	9.78	9.32	9.91

dictability of the structure-activity equation system. Note that prediction refers to the ability of a function to correctly represent data which were not members of the training set, that is, which were excluded from the derivation of the mathematical equation. Hence, the smallness of the least eigenvalue indicates that the three regressors are nearly multicollinear, and warns us that the coefficients of a regression matrix are unreliable and will be inflated, especially, with respect to the prediction of drug effects.

To avoid this problem, the significance of the eigenvalues is examined. As only λ_1 is significant, $\mathbf{v}_1$ is used for the principal component regression analysis. The model is $\mathbf{Y} = \mathbf{B}_z\mathbf{Z}'$ where $\mathbf{Y} = (Y_1\ Y_2)$, $\mathbf{Z} = \mathbf{VX}'$. In our case, we obtain that $Z' = (Z_1\ 0\ 0)$ because $\mathbf{v}_2$ and $\mathbf{v}_3$ belong to eigenvalues that are nonsignificant at the 5% level. Therefore, written in the univariate form, we get that

$$\underset{}{Z_1} = \underset{(32.54\%)}{0.570451\,X_1} - \underset{(32.97\%)}{0.574201\,X_2} + \underset{(34.49\%)}{0.587265\,X_3}$$

This leads to $\mathbf{B}_z =$

$$\begin{bmatrix} 12.25379 & 1.49430 \\ 3.02931 & 1.56271 \end{bmatrix}$$

where $R_1 = 0.98065$, $R_2 = 0.96631$ are identical with the simple correlation coefficients $r(Y_1, Z_1)$ and $r(Y_2, Z_2)$. Table 2 shows the theoretically calculated values (run 3). The correlations between Z_1 and the regressors are highly significant, we get that $r(Z_1, X_1) = 0.97918$, $r(Z_1, X_2) = -0.96809$, $r(Z_1, X_3) = 0.99763$ (0.1% level or less), and the correlations between the biological responses and the transformed values Z_1 (see above) are also significant at the 0.01% level or less.

Note that there also situations that Z_2 must be also included in the principal component regression analysis. In such cases, Z_1 and Z_2 are not statistically collinear, in general.

Pharmacological conclusions

The following conclusions can be deduced from the statistical analysis: a high inhibitory activity against dihydrofolic acid reductase in microorganisms requires that

– at least two or three substituents with strong hydrogen acceptor properties are introduced into the basic molecule (X_1 takes the value of 2 or 3 in such cases),
- substituents with strong positive responance effects must be introduced with electron releasing properties,
– the sum of the steric constants must be relatively large. In such cases, it seems that orientation and intergroup distance in the drug-receptor complex play a role.

When backward elimination or, respectively, stepwise methods are used, the prediction of drug effects is questionable when multicollinearities exist (for instance, it could be assumed that steric effects do not play a role). This can lead to considerable mistakes in drug prediction. The table collection of chemical substitution constants shows that there is no other substituent among 400 substituents that satisfies **each** of the three requirements suggested above. Therefore, it is statistically impossible to improve the activity, and a novel basic molecule must be created.

References

HANSCH, C. [1971]. Quantitative Structure-Activity Relationships in Drug Design. Drug Design, Vol. 1. Academic Press, New York, Ariens, E. J. (Ed.) 271–342.

HANSCH, C., FUKUNAGA, J. Y., JOW, P. Y. C., and HYNES, J. B. [1977]. Quantitative Structure-Activity Relationships of Antimalarials and Dihydrofolate Reductase Inhibition by Quinazolines and 5-Substituted Benzyl-2,4-diaminopyrimidines. J. Med. Chem. *20*, 96–102.

KENDALL, M. [1976]. Some Notes on Statistical Problems Likely to Arise in the Analysis of WFS Surveys. Tech. Bull. No. 441, International Statistical Institute, Voorburg, The Hague.

MAGER, P. P. [1976]. Ein quantitatives Modell der Pharmakochemie. Sci. Pharm. *44*, 40–58, 143-158.

MAGER, P. P. [1979, 1980]. The Masca Model Applied in Drug Design. Drug Design, Vol. 9 & 10. Academic Press, New York, Ariens, E. J. (Ed.).

Anschrift des Verfassers: OA Dr. Peter Mager, Bereich Pharmakologie der Sektion Pharmazie der Fakultät für Naturwissenschaften, Schenkendorfstraße 15, DDR-703 Leipzig

EDV in Medizin und Biologie **10** (3), S. 86–94, ISSN 0300-8282

Experimental Design in Plant Breeding Computer Evaluation of Sugar Beet Varieties

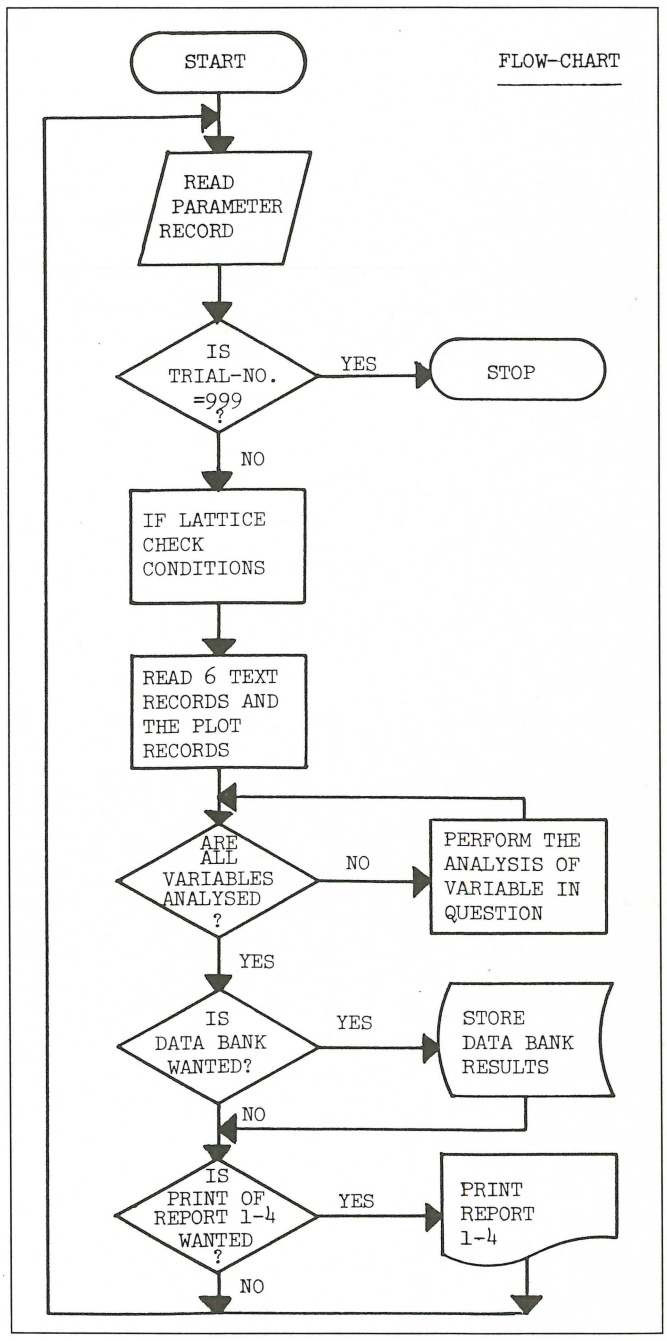
Flemming Yndgaard*

Summary

A system of computer programs (FORTRAN), developed for selection and discarding of breeding material, for balanced and partially balanced lattice designs and for randomized blocks designs is presented. By combining an F-test with a test of »least significant difference«, the program system identifies the probable »winners« and »losers« in a yield trial. The principles of the program system are shown in a flow chart. The specific examples of the use of the system have been drawn from sugar beet breeding. In its present version the system is able to analyse 12 characteristics independently. It is possible to store the results from the analysis in a data bank.

Zusammenfassung

Es wird ein System von Computer-Programmen (FORTRAN) vorgestellt, das für Selektion und Kassieren von Zuchtmaterial, für balancierte und teilweise balancierte Gitteranlagen und für Blockanlagen entwickelt ist. Durch das Kombinieren eines F-Tests mit einem Test von »Grenzdifferenz« identifiziert das Programm-System die wahrscheinlichen »Gewinner« und »Verlierer« eines Leistungsversuches. Die Prinzipien des Programm-Systems sind in einem Flußdiagramm gezeigt. Die spezifischen Beispiele der Anwendung des Systems sind von der Zuckerrübenzüchtung genommen. Das System kann in seiner jetzigen Form 12 Kennzeichnungen unabhängig voneinander analysieren. Es ist möglich, die Ergebnisse der Analyse in einer Data-Bank zu lagern.



Introduction

To overcome the plant breeders' problem of analyzing a great number of varieties or entries in one trial at the same time keeping the trial error at a reasonably low value, Yates developed the lattice designs in the 1930'ies. The relative efficiencies of different lattice designs have been investigated by LeClerg (1966). He found that only 17 of 676 lattice experiments were less efficient than the randomized complete block design. From a study of 81 lattice experiments Ma and Harrington (1948) found that the percentage gains in efficiency compared with randomized block designs were 28, 60, 63, 98 and 32 for

*) This paper was read at the Third Meeting of the Biometric Section of Eucarpia 3rd to 7th April 1978.

simple, triple, quadruple, balanced lattices, and lattice squares, respectively. It is thus obvious that on the average, the plant breeder will save replications by using lattice designs in his investigations.

Statistical investigation of varieties

The process of statistical analysis of varieties by means of computer in plant breeding logically falls into 3 main procedures:

1. Collection and preparation of data for statistical analysis.
2. Statistical analyses.
3. Presentation of the results from the statistical analysis.

Point 1 will not be discussed.
Point 2 will give an instruction in the application of the computer programs developed.
The presentation will concentrate on point 3.

To facilitate the transfer of the system of statistical programs between different computer installations, integer variables have been used for input/output device numbers. For the same reason all input and output statements have been placed in the main program. The output from the statistical analyses establishes a data-bank. The printout presents the result in 4 »Reports«. They will be presented in the paper. The program system, the record specification and the way to handle the analysis system will also be discussed. The source decks of the FORTRAN programs are available on request. Due to many comments the programlist should be self-explanatory when read in connection with the relevant chapters of COCHRAN & COX (1957). For the theory and calculation methods of lattice designs and randomized block designs, the book mentioned above should also be consulted, as it is beyond the scope of this paper to include them.

The use of standards is optional. As shown by YATES (1939) it is legitimate to calculate a lattice design as randomized blocks design. The programs check if the conditions for »lattice« calculation are fullfilled. If not, the program automatically switches from »lattice« to a »randomized blocks« calculation.

Via input the user is also able to switch between »lattice« and »randomized blocks« on the whole or part of the set of variables.

Presentation of the results from the statistical analyses, an example

A sugar beet variety trial is used to demonstrate the results from the statistical analysis by means of a set of 4 reports.

Report 1 gives a summary of the statistical analysis and general information on the trial.

Report 2 is produced when one or more standards are used and gives the relative figures as percentages of the standards.

Report 3 is always produced and gives the relative figures as percentages of the trial mean.

Report 4 reproduces complete information on the analysis of variance for all variables.

The ID number in the top right corner of each sheet identifies the trial. The two first ciphers indicate the year, the next two indicate the country, the next two the region inside the country and the last three indicate the trial number, compare columns 6–14 in the RECORD DEFINITIONS below.

Report 1

As indicated in the second line of report 1, the design code (general) is 2, i.e. we have a lattice design. cfr. Table 1.

The trial consists of 4 blocks or replications and 25 varieties. In the same line it can also be seen that variety number 1, 2 and 3 are used as standards in this trial. In the next line is shown which blocks are present. The statistical summary in the report has been developed to give the breeder the minimum information only.

All variables are calculated independently and it is the variable number which is used to identify the variable by the computer programs. The variable names to the left are only printed to tell what the variable number stands for.

The column for design code indicates the type of calculation used for the single variables. The computer program checks if the necessary conditions for the general design are met. Otherwise the design code is reduced. By means of the number in the column for error code it is then possible to find the explanation for the reduction of the design code compared to the general design code. The explanations are given in the bottom half to the right.

Error code 1 indicates that one or more varieties are missing in the trial. We have made no effort to solve this statistical problem. It means that no statistical tests are performed for the variables which are faulty in that respect. However, beside trial mean the means of the varieties present are given in Report 3 as relative figures in percentage of the trial mean. The design code will be reduced to 0.

Error code 2 indicates that more than 20% of the data are missing from the plots. The actual value of missing data is given in percentage in the column next to the error code column. More than 20% of data missing is so serious a deficiency that no statistical tests are performed. In analogy to error code 1 only trial mean and Report 3 are given for the faulty variables. The design code for those variables will be reduced to 0.

Error code 3 indicates that more than 10% of the data are missing from the plots. It has been suggested by COCHRAN & COX (1957) that lattice designs should be calculated as randomized blocks designs if numerous data are missing. The design code will therefore be reduced from 2 to 1 for lattice designs if data are missing from more than 10% of the plots.

For demonstration of the reduction of design code, 14% and 25% of the data, selected by random, have been discarded for phosphorous and invert sugar respectively. In the original data 1% was missing.

Error codes 4 and 5 are of interest to the statistician only.

Table 1 Program Control and Explanation

Design code	Type of design
1	Randomized blocks
2	Lattice design
9	No calculation wanted
0	Faulty data. (This is set by the computer program; and no statistical analysis is made nor are values for missing data estimated.)

THE BREEDING STATION " M A R I B O " ID NO 770104406

DESIGN CODE (GENERAL)= 2 TOTAL OF BLOCKS= 4 TOTAL OF VARIETIES= 25 STANDARS= 1 2 3 0 0 REPORT 1
BLOCKS PRESENT 1 2 3 4 0 0 0 0 0 0 STATISTICAL SUMMARY

VARIABLE NO	NAME	DESIG. CODE	ERROR CODE	MISSING DATA %	MEAN OF TRIAL	MEAN OF STANDARDS	C.V.	F VALUE	P OF F	REL. PR.	REL. LSD 5%	REL. LSD 1%	REL.STD.LSD 5%	REL.STD.LSD 1%
1	ROOT YIELD..	2	0	0	53.87	51.78	3.6	2.64	.2	112	5.3	7.1	4.3	5.8
2	SUGAR YIELD.	2	0	0	9.39	9.14	3.5	2.33	.5	108	5.0	6.7	4.1	5.5
3	POLARIZATION	2	0	0	17.44	17.63	1.2	3.52	.0	118	1.7	2.2	1.4	1.8
4	POTASSIUM...	2	0	1	927.70	911.65	3.7	6.37	.0	159	5.3	7.1	4.3	5.8
5	SODIUM.....	2	0	1	93.81	79.44	13.5	3.55	.0	124	19.2	25.6	15.7	21.0
6	AMINO-N.....	2	0	1	112.27	111.36	9.1	6.97	.0	150	12.9	17.2	10.6	14.1
7	IMPURITY....	2	0	1	3.77	3.67	4.9	6.00	.0	173	6.9	9.3	5.7	7.6
8	PHOSPHOROUS.	1	3	15	128.94	132.44	4.1	19.30	.0	0	5.8	7.7	4.8	6.3
9	INVERT SUGAR	0	2	26	579.39	.00	.0	.00	.0	0	.0	.0	.0	.0
10	GLUCOSE.....	2	0	1	414.58	414.00	10.3	2.47	.3	101	14.7	19.5	12.1	16.0
11	FREE 1.....	0	0	0	.00	.00	.0	.00	.0	0	.0	.0	.0	.0
12	FREE 2.....	0	0	0	.00	.00	.0	.00	.0	0	.0	.0	.0	.0

VARIABLE NO	CALCULATION CODE USED	COMMENTS TO THE TRIAL
1	1	SOWN ON 1977- 5- 4
2	1	HARVESTED 1977-10-11
3	1	SPRING
4	7	NOT IRRIGATED
5	5	PLOT SIZE SQM: 10.0
6	1	BEET/PLOT(THEOR.) 72
7	1	ROWS SOWN 4
8	1	ROWS HARVESTED 2
9	1	TRIAL PURPOSE
10	1	SPRAYINGS
11		DISCARDED BY RANDOM
12		P(14%) , IS(25%)

ERROR CODE	EXPLANATION
0	NO ERROR DETECTED
1	AT LEAST ONE VARIETY MISSING ALL OVER
2	MORE THAN 20% MISSING DATA
3	MORE THAN 10% MISSING DATA
4	MORE THAN 100 ITTERATIONS NEEDED FOR CALCULATION OF MISSING DATA (R.BLOCKS)
5	MORE THAN 100 ITTERATIONS NEEDED FOR CALCULATION OF MISSING DATA (LATTICE)
6	LATTICE CONDITIONS NOT CORRECT

DATE OF CALCULATION 1977-12- 2 D-BANK CODE 4 COUNTRY DENMARK PLACE ALSTED GD. TRIAL NO 406

Error code 6 indicates that the lattice conditions have not been fulfilled. The design code will be reduced to 1 and the trial calculated as a randomized blocks design.

The column for »mean of trial« gives the trial means of the variables measured in input data units. In the same way the column for »mean of standards« gives the corresponding means of the standards if standards have been introduced in the trial.

The column »C.V.« gives the coefficients of variation.

The volumn »F value« gives the relevant F-ratio for varieties from the analysis of variance. cfr. Report 4 below.

The column »P of F« gives the probability in percent associated with the corresponding F-ratio.

THE BREEDING STATION " M A R I B O " ID NO 770104406

RELATIVE FIGURES IN RELATION TO THE STANDARDS REPORT 2
PAGE 1

VAR	REG.NO.	REP	RR	RS	RP	K	NA	N	IV	P	IS	GL		
1	20000005001720	4	101	102	100	96	94	84=	93-	92=	0	105	0	0
2	30000005001705	4	97	97	99	101	101	92	99	110*	0	94	0	0
3	40000005001705	4	100	99	98=100	102	121*	108*	97		0	101	0	0
4	1150402 59373	4	102	100	97=103	120+	97	103	102		0	94	0	0
5	4150402 68570	4	105+	103	98=104	127*	112+	109*	91=		0	103	0	0
6	5150402 90180	4	100	100	100	90=116+	92	93-	91=		0	97	0	0
7	2150402 90181	4	104	101	97=	98	129*	106	103	93=	0	95	0	0
8	6150402 90280	4	106*	104	98=101	134*	112+	108*	97		0	99	0	0
9	2150402 90281	4	106*	103	97=101	153*	106	106+	95-		0	92	0	0
10	1150402 90380	4	106*	104	98=105+	141*	106	109*	110*		0	99	0	0
11	41504024011150	4	106*	103	98=105+	112	95	102	111*		0	85-	0	0
12	1504024011151	4	101	101	98=103	114	110	106+	111*		0	103	0	0
13	3140802 59313	4	102	100	97=100	100	90	97	98		0	101	0	0
14	3140802 59315	4	104	101	97=104	116+	96	103	94-		0	101	0	0
15	1140802 65570	4	105+	104	99	109*	103	78=	100	96	0	114+	0	0
16	4140802 66800	4	96	96	99	93=101	92	93-	82=		0	117*	0	0
17	5140802 67001	4	106*	104	98=103	116+	98	103	91=		0	101	0	0
18	140802 67070	4	101	100	98=102	105	100	102	92=		0	103	0	0
19	1140802 67070	4	101	99	97=	99	116+	100	101	93=	0	103	0	0
20	3140802 68381	4	107*	105+	99	104	112	103	104	88=	0	94	0	0
21	5140802 71212	4	110*	106*	96=108*	134*	129*	117*	113*		0	119*	0	0
22	7140802 71502	4	108*	106*	98=104	131*	110	109*	101		0	99	0	0
23	3140802 81305	4	104	105+	101	93=	95	84=	91=	80=	0	97	0	0
24	140802 88500	4	107*	104	97=104	135*	100	105	103		0	96	0	0
25	3140802 863370	4	105+	104	98=	94=	131*	86-	95	94-	0	85-	0	0

DATE OF CALCULATION 1977-12- 2 D-BANK CODE 4 COUNTRY DENMARK PLACE ALSTED GD. TRIAL NO 406

The column »Rel. Pr.« gives the relative precision. This is only used when the lattice calculation has been performed. It is a measure of gain in accuracy using lattice design compared to randomized blocks design. The latter is given the value 100.

The two columns »rel. LSD 5% and 1%« give the ordinary »least significant difference« (LSD in the sequel) values for 5% level of significance and 1% level of significance in percentages of the trial mean.

The last two columns »Rel. Std. LSD 5% 1%« are used for comparison of a variety mean to the mean of the standards. The values will always be less than or equal to the corresponding ordinary LSD values, the reason being that the accuracy of the mean of the standards will increase when the number of standards increases.

Pairwise multiple comparison of differences between variety means using LSD ought to be used only if the F-ratio is significant, cfr. CARMER & SWANSON (1973)

In addition to the statistical summary Report 1 contains information about the calculation codes used for preparing input and comments to the trial taken from 6 text records in input stream.

Report 2

This report is produced when standards are used and shows the results of the measured variables in percentage of the mean of the standards.

The first column »VAR« to the left gives the variety number in the trial. The next »REG.NO.« gives the identity of the variety. The column »REP« gives the total number of replications of the variety measured for variable 1.

The next 12 columns present the relative figures for the variables measured.

The columns with relative figures may be followed by a »significance« mark. When varieties as mentioned above by means of an F-test are found to be significantly different, it is legitimate to compare each of them to all the others, to the trial mean and to the mean of the standards.

In report 2 the comparison between the single variety and the mean of the standards has been performed by the computer program if F was significant at the 5% level.

Varieties giving differences numerically greater than the »REL.STD.LSD« by this comparison will be marked with a significance mark to the right of the relative number. The significance marks are given according to the table:

Condition to be fulfilled	Significance Mark	
	5% level	1% level
Relative figure $> \bar{x}_s +$ REL.STD.LSD.	+	*
Relative figure $< \bar{x}_s -$ REL.STD.LSD.	−	=

$\bar{x}_s$ stands for mean of the standards. No adjustment of the LSD value has been made for the comparison when one or both varieties are not present in the full number of replications.

Report 3

Detailed explanations of this report will not be given because it would be exactly the same as for Report 2 except that $\bar{x}_s$ and REL.STD.LSD. in the table for significance marks should be changed to $\bar{\bar{x}}$ and REL.LSD. $\bar{\bar{x}}$ stands for the trial mean and REL.LSD. is the ordinary LSD. It means that the comparisons performed are between the variety means and a hypothetical variety giving the same mean as the trial mean.

THE BREEDING STATION " M A R I B O " ID NO 770104406

RELATIVE FIGURES IN RELATION TO THE TRIAL MEAN REPORT 3 PAGE 1

VAR	REG.NO.	REP	RR	RS	RP	K	NA	N	IV	P	IS	GL		
1	20000005001720	4	98	100	102+	95	80-	84-	91-	95	94	105	0	0
2	30000005001705	4	94-	95	101	100	86	92	97	113*	90	94	0	0
3	40000005001705	4	97	97	100	99	87	121*	105	100	98	101	0	0
4	1150402 59373	4	99	98	99	102	102	97	101	105	93	94	0	0
5	4150402 68570	4	101	101	100	103	108	112	106	94-	104	103	0	0
6	5150402 90180	4	97	98	102+	89=	99	92	91-	94-	99	97	0	0
7	2150402 90181	4	100	99	99	97	110	106	101	96	97	95	0	0
8	6150402 90280	4	102	102	100	100	114	112	105	100	96	99	0	0
9	2150402 90281	4	102	101	99	100	130*	106	104	98	96	92	0	0
10	1150402 90380	4	102	102	100	104	120+	106	106	113*	102	99	0	0
11	41504024011150	4	102	101	100	104	95	95	100	115*	83	85-	0	0
12	1504024011151	4	98	99	100	102	97	110	104	115*	101	103	0	0
13	3140802 59313	4	99	98	99	99	85	90	95	101	98	101	0	0
14	3140802 59315	4	100	99	99	103	99	96	101	97	103	101	0	0
15	1140802 65570	4	101	102	101	106*	88	78=	98	99	119	114	0	0
16	4140802 66800	4	93-	94-	101	92=	86	92	91-	85=	119	117+	0	0
17	5140802 67001	4	102	102	100	102	99	98	101	94-	104	101	0	0
18	140802 67070	4	98	98	100	101	89	100	100	95	96	103	0	0
19	1140802 67070	4	98	97	99	98	100	100	99	96	100	103	0	0
20	3140802 68381	4	103	103	101	103	95	103	102	91=	90	94	0	0
21	5140802 71212	4	106+	104	98-	107+	114	128*	114*	117*	120	119+	0	0
22	7140802 71502	4	104	104	100	103	111	110	106	104	101	99	0	0
23	3140802 81305	4	100	103	103*	92=	81	84-	89=	83=	100	97	0	0
24	140802 88500	4	103	102	99	103	115	100	103	106+	95	98	0	0
25	3140802 863370	4	101	102	100	93-	111	86-	93-	97	82	85-	0	0

DATE OF CALCULATION 1977-12- 2 D-BANK CODE 4 COUNTRY DENMARK PLACE ALSTED GD. TRIAL NO 406

THE BREEDING STATION " M A R I B O " ID NO 770104406
ANALYSIS OF VARIANCE REPORT 4 PAGE 1

VARIABLE 1 ROOT YIELD..

	DF	S S	M S	F	P
TOTAL	99	841.65			
BLOCKS	3	226.47	75.49	15.99	.0
VARIETIES UNADJUSTED	24	275.20	11.47	2.43	.2
BLOCKS DESIGN ERROR	72	338.99	4.72		
VARIETIES ADJUSTED	24	238.35	9.93	2.64	.2
MINIBLOCKS ADJUSTED	16	129.03	8.06		
INTRA MINIBL. ERROR	56	210.96	3.77		
VARIETIES PLOTADJUST	24	270.82	11.28	2.68	.2
EFFECTIVE ERROR	56		4.21		

VARIABLE 2 SUGAR YIELD.

	DF	S S	M S	F	P
TOTAL	99	16.97			
BLOCKS	3	4.23	1.41	11.65	.0
VARIETIES UNADJUSTED	24	6.02	.25	2.07	1.0
BLOCKS DESIGN ERROR	72	8.72	.12		
VARIETIES ADJUSTED	24	5.65	.24	2.33	.5
MINIBLOCKS ADJUSTED	16	3.07	.19		
INTRA MINIBL. ERROR	56	5.65	.10		
VARIETIES PLOTADJUST	24	6.26	.26	2.34	.5
EFFECTIVE ERROR	56		.11		

VARIABLE 3 POLARIZATION

	DF	S S	M S	F	P
TOTAL	99	9.26			
BLOCKS	3	1.43	.48	9.01	.0
VARIETIES UNADJUSTED	24	4.04	.17	3.18	.0
BLOCKS DESIGN ERROR	72	3.81	.05		
VARIETIES ADJUSTED	24	3.34	.14	3.52	.0
MINIBLOCKS ADJUSTED	16	1.60	.10		
INTRA MINIBL. ERROR	56	2.22	.04		
VARIETIES PLOTADJUST	24	3.77	.16	3.50	.0
EFFECTIVE ERROR	56		.04		

VARIABLE 4 POTASSIUM...

	DF	S S	M S	F	P
TOTAL	99	465669.00			
BLOCKS	3	113765.00	37921.67	19.32	.0
VARIETIES UNADJUSTED	24	212598.00	8858.25	4.51	.0
BLOCKS DESIGN ERROR	71	139326.00	1962.34		
VARIETIES ADJUSTED	24	156720.00	6655.03	6.37	.0
MINIBLOCKS ADJUSTED	16	81833.62	5114.61		
INTRA MINIBL. ERROR	55	57402.16	1045.31		
VARIETIES PLOTADJUST	24	185992.00	7749.67	6.30	.0
EFFECTIVE ERROR	55		1230.13		

VARIABLE 5 SODIUM.....

	DF	S S	M S	F	P
TOTAL	96	33537.70			
BLOCKS	3	4304.80	1431.60	7.07	.1
VARIETIES UNADJUSTED	24	14863.12	619.30	3.06	.0
BLOCKS DESIGN ERROR	71	14377.78	202.53		
VARIETIES ADJUSTED	24	12097.32	504.06	3.55	.0
MINIBLOCKS ADJUSTED	16	6566.82	410.43		
INTRA MINIBL. ERROR	55	7812.96	142.05		
VARIETIES PLOTADJUST	24	14446.10	601.92	3.70	.0
EFFECTIVE ERROR	55		162.70		

VARIABLE 6 AMINO-N.....

	DF	S S	M S	F	P
TOTAL	96	34666.62			
BLOCKS	3	2570.81	856.94	5.41	.2
VARIETIES UNADJUSTED	24	20659.12	860.13	5.49	.0
BLOCKS DESIGN ERROR	71	11236.69	158.26		
VARIETIES ADJUSTED	24	14978.62	624.11	6.97	.0
MINIBLOCKS ADJUSTED	16	6314.60	394.66		
INTRA MINIBL. ERROR	55	4922.98	89.46		
VARIETIES PLOTADJUST	24	17006.19	708.59	6.76	.0
EFFECTIVE ERROR	55		104.87		

DATE OF CALCULATION 1977-12- 2 D-BANK CODE 4 COUNTRY DENMARK PLACE ALSTED GD. TRIAL NO 406

Report 4

As mentioned above all variables are analysed independently.

The analysis of variance is also given in exactly the same way for all variables measured.

The variables are identified by the statistical programs by means of their number.

The five columns headed DF, S S, M S, F and P give the degrees of freedom, sum of squares, mean squares, F values and probability in percent associated with the corresponding F value, respectively.

THE BREEDING STATION " M A R I B O " ID NO 770104406
ANALYSIS OF VARIANCE REPORT 4 PAGE 2

VARIABLE 7 IMPURITY....

	DF	S S	M S	F	P
TOTAL	98	12.45			
BLOCKS	3	1.97	.66	10.95	.0
VARIETIES UNADJUSTED	24	6.21	.26	4.31	.0
BLOCKS DESIGN ERROR	71	4.26	.06		
VARIETIES ADJUSTED	24	4.21	.18	6.00	.0
MINIBLOCKS ADJUSTED	16	2.65	.17		
INTRA MINIBL. ERROR	55	1.61	.03		
VARIETIES PLOTADJUST	24	4.86	.20	5.85	.0
EFFECTIVE ERROR	55		.03		

VARIABLE 8 PHOSPHOROUS.

	DF	S S	M S	F	P
TOTAL	84	15286.19			
BLOCKS	3	508.31	169.44	5.96	.2
VARIETIES UNADJUSTED	24	13156.73	548.26	19.30	.0
BLOCKS DESIGN ERROR	57	1619.14	28.41		
VARIETIES ADJUSTED	0	.00	.00	.00	.0
MINIBLOCKS ADJUSTED	0	.00	.00		
INTRA MINIBL. ERROR	0	.00	.00		
VARIETIES PLOTADJUST	0	.00	.00	.00	.0
EFFECTIVE ERROR	0		.00		

VARIABLE 9 INVERT SUGAR

	DF	S S	M S	F	P
TOTAL	0	.00			
BLOCKS	0	.00	.00	.00	.0
VARIETIES UNADJUSTED	0	.00	.00	.00	.0
BLOCKS DESIGN ERROR	0	.00	.00		
VARIETIES ADJUSTED	0	.00	.00	.00	.0
MINIBLOCKS ADJUSTED	0	.00	.00		
INTRA MINIBL. ERROR	0	.00	.00		
VARIETIES PLOTADJUST	0	.00	.00	.00	.0
EFFECTIVE ERROR	0		.00		

VARIABLE 10 GLUCOSE.....

	DF	S S	M S	F	P
TOTAL	98	248884.25			
BLOCKS	3	11301.50	3767.17	2.01	11.9
VARIETIES UNADJUSTED	24	104327.50	4346.86	2.32	.4
BLOCKS DESIGN ERROR	71	133255.37	1876.85		
VARIETIES ADJUSTED	24	104674.37	4369.76	2.47	.3
MINIBLOCKS ADJUSTED	16	36027.26	2251.71		
INTRA MINIBL. ERROR	55	97227.66	1767.76		
VARIETIES PLOTADJUST	24	108430.25	4517.93	2.44	.3
EFFECTIVE ERROR	55		1852.21		

VARIABLE 11 FREE 1.....

	DF	S S	M S	F	P
TOTAL	0	.00			
BLOCKS	0	.00	.00	.00	.0
VARIETIES UNADJUSTED	0	.00	.00	.00	.0
BLOCKS DESIGN ERROR	0	.00	.00		
VARIETIES ADJUSTED	0	.00	.00	.00	.0
MINIBLOCKS ADJUSTED	0	.00	.00		
INTRA MINIBL. ERROR	0	.00	.00		
VARIETIES PLOTADJUST	0	.00	.00	.00	.0
EFFECTIVE ERROR	0		.00		

VARIABLE 12 FREE 2.....

	DF	S S	M S	F	P
TOTAL	0	.00			
BLOCKS	0	.00	.00	.00	.0
VARIETIES UNADJUSTED	0	.00	.00	.00	.0
BLOCKS DESIGN ERROR	0	.00	.00		
VARIETIES ADJUSTED	0	.00	.00	.00	.0
MINIBLOCKS ADJUSTED	0	.00	.00		
INTRA MINIBL. ERROR	0	.00	.00		
VARIETIES PLOTADJUST	0	.00	.00	.00	.0
EFFECTIVE ERROR	0		.00		

DATE OF CALCULATION 1977-12- 2 D-BANK CODE 4 COUNTRY DENMARK PLACE ALSTED GD. TRIAL NO 406

Guidance of the program system

The principles of the program system are shown in the flow chart.

The different types of records used by the statistical programs are identified by means of the contents of columns 1–5.

Columns 1-2 always contain 50, the code for the breeding station »MARIBO« in the computer center.

Column 3 contains the subject code which defines the type of investigation as shown in Table 2.

Columns 4–5 define the type of input or output as shown in Table 3. The character b is used for blank.

The preparation of input data includes preparation of a record of parameters, controlling the statistical programs; the record is shown in RECORD DEFINITION 1.

Columns 6–20 define the trial.

Column 21 contains the d-bank code which must be 4, 5 or 6 if the results are going to be stored in the d-bank,

Column 22 contains the printout code which must be greater than zero if the 4 reports are wanted.

Columns 23–25 define the codes for special analyses. Default value is 0.

Column 26 defines the number of varieties used for »standard«.

Columns 27–41 define the variety numbers for standards.

The maximum number of the system is 12 observations per plot; if less than 12 variables are measured, the number of the last one to be calculated must be written in columns 42–43.

Columns 44–103 contain information controlling the calculation of the 12 single variables. Because they are all calculated independently of each other it is sufficient to look at columns 44–48.

The design code (general) written in column 15 defines the type of design as shown in Table 1. As shown by YATES (1939) it is legitimate to calculate a lattice design as the less complex randomized blocks design. In the present system the user has the liberty to decrease the complexity for one or more variables, i.e. even if column 15 is equal to 2 it is allowed to write 1 in column 44. To make the system independent of the programs producing input data and handling the d-bank, in our case written in COBOL, and the statistical programs written in FORTRAN, all inputs and outputs are transformed into integer formats. 10 raised to the power given in columns 45–46 multiplied by the input data will give the correct unity of input data. E.g. the number 53 is transformed to 530 by multiplication by 10^1 and 5.3 by multiplication by 10^{-1}.

Columns 47–48 contain the value to which the observed results must be compared in order to find the plots from which data are missing. Normally this check value is zero, but for field emergence, for instance, zero could be an observed value. The observed result 0 of a not emerging variety must not be replaced by a higher value by calculating missing data for the plot.

Check value for field emergence is therefore -1. In other words an observation has to be greater than the check value. Otherwise it will be considered as missing data. The explana-

Table 2 Program Control and Explanation

Subject code	Type of investigation
1	Not used
2	Not used
3	Yield Trial
4	Bolter Trial
5	Biochemistry
6	Biology
7	Field Emergence Trial
8	Agricultural Trial
9	Seed Processing

Record definition 1

Column	
1–2	50
3	subject code
4–5	blanks
6–7	trial identification, year
8–9	trial identification, country
10–11	trial identification, region inside the country
12–14	trial identification, trial number
15	trial identification, design code (general)
16–17	trial identification, total of blocks
18–20	trial identification, total of varieties
21	d-bank code
22	print code
23–25	codes for special analyses
26	standards, total of standards
27–29	standards, 1
30–32	standards, 2
33–35	standards, 3
36–38	standards, 4
39–41	standards, 5
42–43	last variable to calculate
44	variable 1, design code
45–46	variable 1, exponent of 10
47–48	variable 1, check value
49	variable 2, design code
50–51	variable 2, exponent of 10
52–53	variable 2, check value
54	variable 3, design code
55–56	variable 3, exponent of 10
57–58	variable 3, check value
59	variable 4, design code
60–61	variable 4, exponent of 10
62–63	variable 4, check value
64	variable 5, design code
65–66	variable 5, exponent of 10
67–68	variable 5, check value
69	variable 6, design code
70–71	variable 6, exponent of 10
72–73	variable 6, check value
74	variable 7, design code
75–76	variable 7, exponent of 10
77–78	variable 7, check value
79	variable 8, design code
80–81	variable 8, exponent of 10
82–83	variable 8, check value
84	variable 9, design code
85–86	variable 9, exponent of 10
87–88	variable 9, check value
89	variable 10, design code
90–91	variable 10, exponent of 10
92–93	variable 10, check value
94	variable 11, design code
95–96	variable 11, exponent of 10
97–98	variable 11, check value
99	variable 12, design code
100–101	variable 12, exponent of 10
102–103	variable 12, check value

Record Definition 2

Column	
1–2	50
3	subject code
4–5	00 or 50
6–7	trial identification, year
8–9	trial identification, country
10–11	trial identification, region inside the country
12–14	trial identification, trial number
15	trial identification, design code (general)
16–17	trial identification, total of blocks
18-20	trial identification, total of varieties
21–23	variety identification, number in trial
24–37	variety identification, registration number
38–39	plot identification, block in field
40–41	plot identification, row within block
42–43	plot identification, column within block
44–46	plot identification, plot no in field
47	discard code
48–52	variable 1 (decimal point defined by parameter-record)
53–57	variable 2 (decimal point defined by parameter-record)
58–62	variable 3 (decimal point defined by parameter-record)
63–67	variable 4 (decimal point defined by parameter-record)
68–72	variable 5 (decimal point defined by parameter-record)
73–77	variable 6 (decimal point defined by parameter-record)
78–82	variable 7 (decimal point defined by parameter-record)
83–87	variable 8 (decimal point defined by parameter-record)
88–92	variable 9 (decimal point defined by parameter-record)
93–97	variable 10 (decimal point defined by parameterrecord
98–102	variable 11 (decimal point defined by parameterrecord)
103–107	variable 12 (decimal point defined by parameterrecord)

tions for all the other 11 variables are analogue. Observed results from each plot are combined into one record, shown in RECORD DEFINITION 2. Columns 1–20 are equal to the parameter record except for columns 4–5 which are 00.

Columns 21–37 define the variety by its number in the trial 21–23 and its registration number 24–37.

Columns 38–46 define the plot. The plot records must be sorted in increasing order with respect to the plot number. This is necessary because the computer programs automatically calculate to which block and miniblock a plot belongs. Block is used as name for complete block equal to replication and miniblock is used as name for incomplete block.

Column 47 is assigned the value 1 if the plot is discarded, but the record must be present.

The observed results for the variables present are written in the respective columns. A maximum of 5 digits has been found suffice for all variables.

For a lattice design the adjustments are made at plot level and the adjusted plot results are kept in data-bank records for further analysis. Their number is 50 in the case of adjusted plot results in columns 4–5. The other identifications are equal to the records for the observed results. RECORD DEFINITION 3 shows the records used for variety totals. Columns 1–37 are equal to the plot record in RECORD DEFINITION 2 except for columns 4–5 which are 51 or 52 as indicated in table 3.

Column 38 is equal to 1 for records containing information on the variables 1–6 and equal to 2 for records containing information on the variables 7–12.

Column 41 is equal to the design code used for the calculation of variable 1.

Columns 42–43 contain the total number of replications of the variety in question.

Columns 44–49 contain the sum of the replications including possible estimates for missing data. Variables 2–12 are treated similarly.

In the analyses of a series of experiments the sums of variety are used in some cases, in other cases, the means of variety COCHRAN & COX (1957). By storing the variety sums in the data bank both possibilities are available without rounding errors.

Results from the analysis of variance are written in records as RECORD DEFINITION 4.

Columns 1–20 are equal to the plot record except for columns 4-5 which are 53. Columns 21–25 contain information on the variable, number (21–22), record number within variable (23–24) and the design code used for the calculation (25). As shown in table 4, nine records are written for each variable. Columns 26-28 contain the relevant number of degrees of freedom. The relevant sum of squares and mean squares are written in columns 29–46 and 47–64 respectively. If an F-test is relevant, columns 65-69 contain the F-value and columns 70–73 contain the probability in percent associated with the F-ratio.

In addition to the results from analyses of variance further statistics are given as indicated in RECORD DEFINITION 5. Columns 1–20 are equal to the plot record except for columns 4–5 which are 54. Columns 21–27 contain information of the variable, its number (21–22), design code used for the calcula-

Table 3 Program Control and Explanation

Column 4–5	Type of input or output
bb	Parameter Record
00	Observed plot results
50	Adjusted plot results (from lattice design calculation)
51	Variety totals from randomized block calculation
52	Variety totals from lattice calculation
53	Analysis of variance
54	Some other statistics

Table 4 Program Control and Explanation

Record number within variable	Source of variation
1	Total
2	Blocks
3	Varieties unadjusted
4	Block design error
5	Varieties adjusted
6	Miniblock adjusted
7	Intra miniblock error
8	Varieties plot-adjusted
9	Effective error

Record definition 3, part 1

Column	
1–2	50
3	subject code
4–5	51 or 52
6–7	trial identification, year
8–9	trial identification, country
10–11	trial identification, region inside the country
12–14	trial identification, trial number
15	trial identification, design code
16–17	trial identification, total of blocks
18–20	trial identification, total of varieties
21–23	variety identification, number in trial
24–37	variety identification, registration number
38	record number 1
39–40	last variable calculated
41	variable 1, design code
42–43	variable 1, total rep. of variety
44–49	variable 1, total of variety
50	variable 2, design code
51–52	variable 2, total rep. of variety
53–58	variable 2, total of variety
59	variable 3, design code
60–61	variable 3, total rep. of variety
62–67	variable 3, total of variety
68	variable 4, design code
69–70	variable 4, total rep. of variety
71–76	variable 4, total of variety
77	variable 5, design code
78–79	variable 5, total rep. of variety
80–85	variable 5, total of variety
86	variable 6, design code
87–88	variable 6, total rep. of variety
89–94	variable 6, total of variety

Record definition 3, part 2

Column	
1–2	50
3	subject code
4–5	51 or 52
6–7	trial identification, year
8–9	trial identification, country
10–11	trial identification, region inside the country
12–14	trial identification, trial number
15	trial identification, design code (general)
16–17	trial identification, total of blocks
18–20	trial identification, total of varieties
21–23	variety identification, number in trial
24–37	variety identification, registration number
38	record number 2
39–40	last variable calculated
41	variable 7, design code
42–43	variable 7 total rep. of variety
44–49	variable 7, total of variety
50	variable 8, design code
51–52	variable 8, total rep. of variety
53–58	variable 8, total of variety
59	variable 9, design code
60–61	variable 9, total rep. of variety
62–67	variable 9, total of variety
68	variable 10, design code

Record definition 3, part 2

Column	
69–70	variable 10, total rep. of variety
71–76	variale 10, total of variety
77	variable 11, design code
78–79	variable 11, total rep. of variety
80–85	variable 11, total of variety
86	variable 12, design code
87–88	variable 12, total rep. of variety
89–94	variable 12, total of variety

Record definition 4

Colum	
1–2	50
3	subject code
4–5	53
6–7	trial identification, year
8–9	trial identification, country
10–11	trial identification, region inside the country
12–14	trial identification, trial number
15	trial identification, design code (general)
16–17	trial identification, total of blocks
18–20	trial identification, total of varieties
21–22	variable, number
23–24	variable, record number
25	variable, design code
26–28	degrees of freedom
29–46	sum of squares (4 dec)
47–64	mean square (4 dec)
65–69	f value (2 dec)
70–73	probability of f (1 dec)

Record definition 5

Column	
1–2	50
3	subject code
4–5	54
6–7	trial identification, year
8–9	trial identification, country
10–11	trial identification, region inside the country
12–14	trial identification, trial number
15	trial identification, design code
16–17	trial identification, total of blocks
18–20	trial identification, total of varieties
21–22	variable, number
23–24	variable
25	variable, design code
26–27	variable, error code
28–30	missing data in percent
31–48	mean of trial (4 dec)
49–66	mean of standards (4 dec)
67–70	coefficient of variation (1 dec)
71–74	relative precision
75–78	relative lsd 15% (1 dec)
79-82	relative lsd 1% (1dec)
83–86	relative std. lsd 5% (1 dec)
87–90	relative std. lsd 1% (1 dec)

tion (25) and code for error detected by the statistical programs (26–27); columns 23–24 are not used. Total of missing plot results in percentage of the whole trial is written in columns 28–30. The trial mean is written in columns 31–48. If one or more varieties are used as standard in the trial their mean is written in columns 49–66.

Columns 67–70 contain the coefficient of variation. Relative precision is written in columns 71–74. Columns 75–90 contain LSD values in percentage of the trial mean. Columns 75–78 and 79–82 are the ordinary LSD values for the 5% and the 1% levels respectively.

Columns 83–86 and 87–90 contain the corresponding LSD values for comparison of the mean of one variety with the mean of the standards.

For the statistics, with a fixed number of decimals, kept in the data bank the number of decimals is indicated in the definition for example mean of trial (4 dec) = 4 decimals.

The parameter record and the 6 text records used for the example are shown on page 25 on the top. On the same page the first and the last 20 plot input records are shown for the same trial.

A complete list of the FORTRAN program system is available on request.

The program system has been developed and checked out on the UNIVAC 1110 at the Copenhagen University Computer Center.

Users employing the program system on other computers may obtain slightly different numerical results because of the different word length of these other computers. If double precision versions of the programs are desired, the C in column 1 should be removed from the double precision statements in the programs.

Research workers planning to implement the system might wish to obtain a punched deck containing all the programs rather than run the risk of introducing errors during repunching from the print-ups. Persons interested in purchasing such a deck should write to the author.

The author is indebted to Mr. Niels Faerch for profitable discussions.

References

CARMER, S. G. and SWANSON, M. R., »An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods«, Journal of the American Statistical Association, Application Section 68, 66–74, 1973.

COCHRAN, W. G. and COX, G. M., »Experimental Designs«, Second Edition, Wiley New York 1957.

LECLERG, E. L., »Significance of Experimental Designs in Plant Breeding«. In: »Plant Breeding«. Edited by K. J. Frey. The Iowa State Universesity Press, Ames, Iowa 1966.

MA, R. H., and HARRINGTON, J. B., »The standard errors of different designs of field experiments at the University of Saskatchewan'. Sci. Agr. 28, 461–474, 1948.

YATES, F., »The recovery of inter-block information in variety trials arranged in three-dimentional lattice«. Ann. Eugenics. 9, 136–156, 1939.

Anschrift des Verfassers: Flemming Yndgaard, Cand. scient., Head of Statistics, The Breeding Station »Maribo«, 4960 Holeby, Denmark

NACHRICHTEN UND BERICHTE

5th Meeting of the International Society of Haematology Hamburg, 26.–31. August 1979
Symposium: Mathematical models of normal and abnormal haemopoietic cell renewal systems

Unter diesem Titel wurde – erstmals in der Geschichte des internationalen Hämatologenkongresses – eine Sitzung über mathematische Modelle abgehalten. Hierbei wurden fünf Vorträge präsentiert.

Im ersten Referat gab WICHMANN (Köln) eine Einführung in die Problematik. Anhand von Beispielen aus dem Bereich der Blutbildung erläuterte er, was ein mathematisches Modell ist, wie es arbeitet und was es leisten kann. Abschließend gab er Kriterien an, die es dem Biologen oder Mediziner erleichtern sollen, auch bei bescheidenen mathematischen Kenntnissen die Qualität und Brauchbarkeit von Modellen zu beurteilen.

Aus der gleichen Arbeitsgruppe stellte LÖFFLER (Köln) ein Stammzellmodell vor. Es versucht, die große Fülle experimenteller Daten (akute und chronische Bestrahlung, Anämie, Hypertransfusion, Hypoxie, Hyperoxie, Erythropeotingabe u. a.) zu reproduzieren und widerspruchsfrei zu interpretieren. Letzteres gelingt bei 90% der verfügbaren Daten.

LORD (Manchester) stellte neuere experimentelle Ergebnisse zur Wirkung von Inhibitoren und Stimulatoren im Stammzellbereich sowie zum Einfluß von Mehrfachtransplantationen vor, die er mathematisch analysierte. Ferner beschrieb er die räumliche Anordnung der Stammzellen im Knochenmark.

PABST (Ulm) diskutierte ein Modell zur Granulopoese bei Hunden. Es ist in der Lage, neben den Normalwerten die Daten zur Leukapherese (Isolierte Entfernung weißer Blutzellen aus der Zirkulation) und zur Gabe von Cyclophosphamid (Zerstörung proliferierender Zellen) zu reproduzieren.

Als letzte Referentin stellte REINCKE (Harvard) ein einfaches Modell vor, welches das Wechselspiel zwischen proliferierenden und ruhenden Stammzellen bei gesteigertem Bedarf an differenzierten Zellen beschreibt.

Ein wichtiges Ziel der Sitzung, nämlich die Sprachbarrieren zwischen Experimentoren und Modellbauern zu verkleinern und gleichzeitig die experimentellen Gruppen auf den Nutzen mathematischer Modelle für ihre eigene Arbeit hinzuweisen, wurde, zumindest teilweise, erreicht. Dies zeigt sich u. a. darin, daß am Rande der Tagung mehrere konkrete Absprachen für eine Zusammenarbeit bei der Planung und Auswertung zukünftiger Experimente getroffen wurden.

Dr. H. E. Wichmann, Med. Universitätsklinik, Jos.-Stelzmann-Straße 9, D-5000 Köln 41

Biometrie – heute und morgen
Interregionales Biometrisches Kolloquium vom 17. bis 20. März 1980 in München

Informationen und Anmeldungen: Dr. W. Koepcke, Institut f. Med. Informationsverarbeitung, Statistik und Biomathematik, Marchioninistraße 15, D-8000 München 70

Biometrische Tagung in Budapest 1981

Es ist vorgesehen, zu Beginn der Osterwoche 1981 eine Biometrische Tagung in Budapest zu veranstalten.

Vortragsanmeldungen und Anfragen an: Dr. János Sváb, Universitas Scientiarum Agrariarum, Facultas Scientiae Agrariae, Cathedra Agriculturae Cultusque Herbarum, H-2103 Gödöllö/Hungaria

BUCHBESPRECHUNGEN

SCHNEIDER, B. und RANFT, U.
Simulationsmethoden in der Medizin und Biologie
Medizinische Informatik und Statistik, Bd. 8, 1978, 496 S., DM 55,–
Springer-Verlag, Berlin – Heidelberg – New York
Die Simulation von technischen, wirtschaftlichen, biologischen oder allgemeinen wissenschaftlichen Problemen mit Hilfe von Rechenanlagen stellt im Augenblick ein beachtliches Hilfsmittel der wissenschaftlichen Forschung dar. Besonders im Bereich der Biologie, der Medizin, der Sozialwissenschaft und der Ökonometrie scheint die Computersimulation eine methodische Lücke zu schließen, da hier meist sehr komplexe Systeme exakten Experimenten nur sehr schwer zugänglich sind bzw. geschlossene Lösungen adäquater mathematischer Modelle nicht vorliegen.

Im Herbst 1977 fand zu diesem Thema ein Workshop statt, um einen Überblick über die Aktivitäten vornehmlich im deutschen Sprachraum zu erhalten. Die bei diesem Workshop gehaltenen Referate sind in dem vorliegenden Band unter folgenden Themenbereichen zusammengestellt: Methoden, Evolutionsprozesse, Physiologische Systeme, Zellkinetische Systeme, Optimierung von Diagnose und Therapie, Ökonomische Systeme, Lernprozesse.

Damit liegt eine ausgezeichnete Übersicht über dieses Fachgebiet vor, das noch recht interessante Anwendungen der Computertechnik erwarten läßt. Ge.

SCHREINER, H.
Computer Cartoons
1979, 96 Seiten mit 102 Abb., DM 16,80
Verlagsges. Rudolf Müller, Köln
Der »tierische Ernst« gedeiht auch in der Umgebung der Computer. Mit manchmal nur wenigen Strichen versteht es der Verfasser, typische Situationen zu charakterisieren. – Ein Buch zur Entspannung, nicht zum Lesen. Ge.

SCHAFFLAND, H. J. und WILTFANG, N.
Bundesdatenschutzgesetz (BDSG)
– Ergänzbarer Kommentar nebst einschlägigen Rechtsvorschriften
Ergänzbare Ausgabe, einschl. 3. Lieferung, 374 S., DM 38,–, zuzügl. Ordner DM 9,80. E. Schmidt Verlag, Berlin – Bielefeld – München
Das Bundesdatenschutzgesetz ist für alle Betroffenen noch recht neu. Daher ist es zu begrüßen, daß hier neben dem Gesetzestext Textauszüge der das BDSG tangierenden Gesetze und Verordnungen zusammengestellt und durch Kommentare erläutert wurden. Durch die Anlage als ergänzbares Werk wird sichergestellt, daß Änderungen leicht eingefügt werden können.

Der Kommentar zeichnet sich durch eine klare Gliederung und eine Vielzahl praktischer Beispiele aus, die auch für Nichtjuristen gut verständlich sind. Von besonderem Wert sind die mitgeteilten Formulierungsvorschläge etwa für die Verpflichtung von Mitarbeitern, für Benachrichtigungen sowie als weitere Arbeitsunterlage der abgedruckte »Leitfaden für den Datenschutzbeauftragten«. – Insgesamt ein nützlicher Ratgeber für diese neue Materie. Ge.

LANGE, H.-J., MICHAELIS, J. und UBERLA, K. (Hrsg.)
15 Jahre Medizinische Statistik und Dokumentation
Aspekte eines Fachgebietes
Medizinische Informatik und Statistik Bd. 9, 1978, 205 S., DM 30,–
Springer-Verlag, Berlin – Heidelberg – New York
Am 30. Januar 1978 vollendete Herr Prof. Dr. Dr. Siegfried Koller sein 70. Lebensjahr. Aus diesem Anlaß fand in Mainz eine Festvorlesung und ein Symposium statt. Die hierbei gehaltenen Vorträge und Referate sind in dem vorliegenden Band zusammengestellt.

Diese Veranstaltungen fanden fast genau 15 Jahre nach dem Beginn des Aufbaues des Mainzer Instituts für Medizinische Statistik und Dokumentation durch Herrn Professor Koller statt. Es war daher Gelegenheit, diese Zeitspanne einmal würdigend zu überblicken. – Die Vielzahl und die Verschiedenartigkeit der Referate veranschaulichen sehr eindrucksvoll die Impulse, die für das neue Fachgebiet in dieser Zeit von Mainz ausgegangen sind. Ge.

SCHACH, S. und SCHÄFER, TH.
Regressions- und Varianzanalyse
1978, 262 S., DM 29,–
Springer-Verlag, Berlin – Heidelberg – New York
Die Regressions- und die Varianzanalyse sind nicht zuletzt durch die Bereitstellung entsprechender Computerprogramme zu sehr verbreiteten statistischen Auswertungsverfahren geworden. Dennoch ergeben sich bei den Anwendungen immer wieder Probleme, die häufig mit den jeweiligen Voraussetzungen zusammenhängen.

Hier setzen die Verfasser ein, die einerseits den Anwendern einen Einblick in die mathematisch-theoretische Fundierung dieser Verfahren geben wollen, andererseits aber auch den Mathematikern und Statistikern mit einem Ausbildungsschwerpunkt auf dem Gebiet der Stochastik einen Überblick über eine Klasse wichtiger Verfahren liefern wollen. – Dieser doppelten Aufgabenstellung wird die Darstellung voll gerecht, wenn auch die Anwender mit geringen mathematischen Vorkenntnissen vielleicht einige Schwierigkeiten haben werden. Umsomehr sollte aber die saubere, klare Darstellung die Mathematiker und Statistiker ansprechen, die bereitgestellten statistischen Verfahren korrekt einzusetzen und die Anwender entsprechend zu beraten. Ge.

BORCHARDT, K.
Die wissenschaftliche Literatur
– Medium wissenschaftlichen Fortschritts
1978, 29 S.
Arbeitsgemeinschaft wissenschaftlicher Literatur e. V., Stuttgart
Die vorliegende Broschüre enthält eine überarbeitete Fassung eines Vortrags des Verfassers. Angesprochen werden darin eine Fülle relevanter Themen im Zusammenhang mit wissenschaftlichen Publikationen, die sowohl Produzenten als auch Konsumenten wissenschaftlicher Literatur, d. h. eigentlich alle Wissenschaftler zum Nachdenken über einen wesentlichen Teil ihrer Tätigkeit anregen sollten. Ge.

PRECHT, M. und VOIT, K.
Mathematik für Nichtmathematiker – Teil 1
1979, 135 S., DM 16,80
R. Oldenbourg Verlag, München – Wien
In diesem ersten Teil einer »Mathematik für Nichtmathematiker« werden insbesondere Grundbegriffe, Vektorrechnung, Matrizenrechnung und lineare Gleichungssysteme behandelt. Die Darstellung ist aus Vorlesungen für Studierende der Agrarwissenschaften, des Gartenbaus, des Brauwesens, der Lebensmitteltechnologie sowie der Ökotrophologie hervorgegangen. Dabei wird versucht, an Beispielen die Ma-

thematik zu erklären, ohne aber die erforderliche Exaktheit dabei zu vergessen. – Die Abgrenzung des Stoffes bereitet auch hier den Verfassern einige Schwierigkeiten. Man kann darüber diskutieren, welche Teile der Mathematik für welche Ausbildungsgänge notwendig sind. Eine solche Diskussion wird wohl immer mit einem Kompromiß enden müssen. Aus dieser Sicht liegt hier insbesondere für Studierende ein brauchbares, vorlesungsbegleitendes Buch vor. Ge.

SCHUSTER, W. und VON LOCHOW, J.
Anlage und Auswertung von Feldversuchen
2., erw. Aufl., 1979, 240 S., DM 40,–
DLG-Verlags GmbH, Frankfurt a. M.
Der Feldversuch liefert der landwirtschaftlichen Praxis und den Landbauwissenschaften Informationen, um die Leistungen von Sorten und die Wirkungen von Anbaumaßnahmen (Düngung, Pflanzenschutz, Ackerbau) unter den verschiedensten Anbaubedingungen beurteilen zu können; er ist durch nichts zu ersetzen. Das Buch »Anlage und Auswertung von Feldversuchen« gibt einen guten Einstieg in die Probleme und ist speziell für diejenigen, die die Planung, technische Durchführung und Verrechnung der Versuche selbst vornehmen müssen, ein unentbehrlicher Leitfaden. Die verständliche Art der Darstellung erleichtert dieses Vorhaben.

Die wesentlich erweiterte Neuauflage enthält im ersten Abschnitt die Darlegung der klassischen Anlagemethoden und der Methoden der Blockanlage sowie der Gitteranlage und der mehrfaktoriellen Versuchsanstellung.

Im zweiten Abschnitt werden anhand von Rechenbeispielen die Auswertungen für die verschiedenen Versuchsanlagen ausführlich beschrieben. In der Neuauflage wurden diese Anleitungen auf die Zwei- und Dreisatzgitter und die Versuchsserien erweitert. Ferner sind Anregungen für die Interpretation von Versuchsergebnissen enthalten. Aus diesem Grunde werden u. a. auch die Grundlagen der Regressions- und Korrelationsrechnung dargestellt. Hau.

DUCROT, H. et al. (Ed.)
Computer Aid to Drug Therapy and to Drug Monitoring
1978, 449 S., $ 49,–
North-Holland Publ. Comp., Amsterdam–New York
Die vorliegenden Proceedings der IFIP TC-4 Konferenz von 1978 über den Computereinsatz bei der Arzneimitteltherapie und der Arzneimittelüberwachung umfassen ein weites Themenspektrum. Dabei war der für jede Aufgabe eingesetzte Computer vielleicht das Bindeglied zwischen den divergierenden Bereichen.

Es wurden u. a. behandelt: die Überwachung von Arzneimittelwirkungen im Krankenhaus und bei ambulanten Patienten, Arzneimitteldatenbanken, die Verteilung von Arzneimitteln sowie Arzneimittelstatistiken.
Damit werden hier die vielschichtigen Probleme bei der Verwendung von Arzneimitteln angesprochen, bei denen der Computer hilfreich sein kann. Ge.

Datenverarbeitung auf dem Umweltsektor
1978, 149 S., DM 20,80
R. Oldenbourg Verlag, München–Wien
In dem vorliegenden Band sind die Referate einer gleichnamigen Tagung der Gesellschaft für Mathematik und Datenverarbeitung und des Umweltbundesamts zusammengestellt.

Da eine laufende Überwachung vielfältiger Umweltbelastungen, eine wirksame Kontrolle der Einhaltung von Grenzwerten für kritische Immissionen und auch eine umweltbezogene Regional- und Wirtschaftsplanung sich nur mit Hilfe der Datenverarbeitung durchführen lassen, ist es erfreulich, daß hier konkrete Einsatzbeispiele und realistische Planungen einmal ausführlich dargestellt werden. Ge.

LITTMANN, H. E. (Hrsg.)
Handbuch der modernen Datenverarbeitung
Grundwerk, rd. 3000 S. in 5 Plastikordnern, DM 139,–. Jährlich 6 Nachlieferungen, DM 68,–
Forkel-Verlag, Stuttgart – Wiesbaden
Das HMD erscheint in Loseblattform, um so mit der ständig fortschreitenden Entwicklung in der EDV Schritt halten zu können.

Die Schwerpunkte dieses Standardwerkes liegen in der Behandlung organisatorischer Grundlagen der Datenverarbeitung, der Erörterung von Informationssystemen sowie in der Darstellung von Rechtsproblemen und Berufen bzw. Berufsbildern in der EDV. Dabei werden z. T. recht ausführlich (bis zu Blockdiagrammen und Formularen) einzelne Anwendungen behandelt.

Mit Angaben zu neuen Entwicklungen in der Hard- und Software liegt hier ein umfassendes Nachschlagewerk vor, das eigentlich in keiner EDV-Abteilung fehlen sollte. Ge.

RIEDWYL, H.
Schweizer Zahlenlotto – Spiel, Zufall und Gewinn
1979, 63 S., DM 13,50
Verlag Paul Haupt, Bern
Der Autor zeigt in einer recht anschaulichen und klaren Art, welche Schlüsse ein Statistiker aus Lottoergebnissen und Lottotips zu ziehen vermag. Seine Darstellung ist dabei auf der einen Seite eine ausgezeichnete Beispielsammlung für die Anwendung statistischer Techniken und auf der anderen Seite eine Anregung für eine mögliche Änderung des Verhaltens von einigen Spielern. – Eine entsprechende Darstellung auch für das deutsche Zahlenlotto wäre gleichfalls wünschenswert, dürfte aber in der Tendenz kaum wesentliche Abweichungen aufzeigen. Ge.

ENDERLEIN, W.
Projektierung von Anwenderdatennetzen
1979, 287 S., DM 87,–
R. Oldenbourg Verlag, München–Wien
In der praktischen Anwendung werden immer mehr Datennetze benötigt. Hier in jedem Fall die bestmögliche Lösung auch unter Berücksichtigung von wirtschaftlichen Gesichtspunkten zu finden, ist nicht immer leicht.

Es ist daher zu begrüßen, daß es der Autor unternommen und verstanden hat, ausgehend von klaren Definitionen über die Darstellung der notwendigen statistischen Begriffe, z. T. gestützt auf übersichtliche Diagramme, hier eine Darstellung dieses komplexen Gebiets vorzulegen. Die angegebenen technischen Beispiele ermöglichen es, die behandelten Zusammenhänge an konkreten Daten zu erproben. Sie werden darüberhinaus konsequent bis zu den Kostenrechnungen fortgeführt. –

Ein unentbehrliches Buch für jeden, der Datennetze plant oder beurteilen will. Ge.

Datenverarbeitung auf dem Umweltsektor
Vorträge einer gemeinsamen Veranstaltung.
Herausgegeben von der Gesellschaft für Mathematik und Datenverarbeitung mbH Bonn und dem Umweltbundesamt Berlin
1978, 149 S., DM 20,80
R. Oldenbourg Verlag, München–Wien
Ziel der Vortragsveranstaltung war es, das bereits vorliegende Material darzustellen, spezifische EDV-Probleme zu behandeln, Lösungsmöglichkeiten aufzuzeigen und auf erkennbare Forschungsschwerpunkte hinzuweisen.

Interessant ist, wie weit schon durch die kontinuierliche Erfassung von Meßwerten über Rechnernetze Erkenntnisse z. B. über die Belastung der Umwelt gewonnen werden. Darüberhinaus sind natürlich eine Reihe von Projekten erst im Planungsstadium. Sie lassen aber hoffen, daß auch mit Hilfe der EDV der Umweltschutz konsequent betrieben wird. Ge.